Spring 2014

Characterization of a plant gene family expanded in glycine max

Lisa Snoderly-Foster

Department:

CHARACTERIZATION OF A PLANT GENE FAMILY EXPANDED IN *GLYCINE MAX*

by

LISA SNODERLY-FOSTER

A THESIS

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE IN APPLIED AND ENVIRONMENTAL BIOLOGY

2014

Approved by

Ronald Frank, Advisor

Katie Shannon

Dave Westenberg

ABSTRACT

Glycine max, commonly named the cultivated soybean, is one of the oldest and most important food crops in the world. The study of the *G. max* genome provides valuable insight into the molecular mechanisms that govern its reproduction and environmental responsiveness, key factors in maximizing crop yield. Since the complete sequencing of the genome in 2010, the analysis has become faster and easier, especially with the development of numerous web-based, publically accessible bioinformatics tools.

This research effort utilizes these tools to characterize a small, unannotated *G. max* gene family. Although no definitive evidence was uncovered for the production of a functional protein product from these genes, evidence does exist for the transcription of 3 of 5 genes. Through gene model verification, synonymous substitution calculations, structural fold analysis, *cis*-element identification, and comparisons to molecules of known structure, an attempt was made to define the evolutionary history and pinpoint putative function of the conceptually translated amino acid sequences from this family of genes.

ACKNOWLEDGMENTS

I would like to extend thanks to Dr. Ronald Frank for his mentorship. He saw enough potential in me to take me on as a student and to invest a substantial amount of time and energy in personally guiding me through this process. Your wisdom has been invaluable.

I would also like to extend thanks to Dr. Dave Westenberg and Dr. Katie Shannon, members of my thesis committee. Thank you each for allowing me to do a rotation in your lab. I gained valuable experience and was afforded the opportunity to learn techniques that I might not have had a chance to learn otherwise.

Dr. Gayla Olbricht, thank you for taking the time to look into my research and help me determine whether a statistical analysis could be performed on the data I collected.

Finally, I would like to extend gratitude to my family for their support. To my partner Jennifer, without your constant reassurance that this was the right path for me and that the financial hardships have been worth the end gain, I might not have had the resolve to give this my best effort. You have been the foundation of my success. To my parents, thank you for supporting this venture, and for your encouragement and willingness to help in any way possible.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

NOMENCLATURE

Nucleotides

Symbol	Description
A	adenine
G	guanine
T	thymine
C	cytosine
U	uracil
R	A or G
Y	C or T
W	U or A

Amino Acids

Symbol	Description	Symbol	Description
A	Ala/alanine	N	Asn/asparagine
C	Cys/cysteine	P	Pro/proline
D	Asp/aspartic acid	Q	Gln/glutamine
E	Glu/glutamic acid	R	Arg/arginine
F	Phe/phenylalanine	S	Ser/serine
G	Gly/glycine	T	Thr/threonine
H	His/histidine	V	Val/valine
I	Ile/isoleucine	W	Try/tryptophan
K	Lys/lysine	Y	Tyr/Tyrosine
L	Leu/leucine		
M	Met/methionine		

1. INTRODUCTION

1.1. *GLYCINE MAX*

Glycine max (L.) Merr., commonly known as cultivated soybean, is a diploidized tetraploid (2n = 40) plant species [1] with agricultural significance in Eastern North America [2] and Asia [1]. The classification of this herbaceous, annual legume, as reported by the USDA Natural Resource Conservation Service Plants Database, is summarized in Table 1.1 [2]. There are two major legume (family Fabaceae [2]) lineages, Hologalegina and Phaseoloides (*Glycine*). From the Phaseoloid line, two subgenera diverged [3]. *Glycine max* and its wild predecessor, *Glycine soja*, belong to the subgenus *Soja*. These species are capable of hybridizing which has implications on gene flow [1].

Table 1.1. Classification of *Glycine max (L.) Merr.* [2]

Level	Scientific Name	Common Name
Kingdom	Plantae	Plants
Subkingdom	Tracheobionta	Vascular Plants
Superdivision	Spermatophyta	Seed Plants
Division	Magnoliophyta	Flowering Plants
Class	Magnoliopsida	Dicotyledons
Subclass	Rosidae	----------
Order	Fabales	----------
Family	Fabaceae	Pea, Legume
Genus	*Glycine Willd.*	Soybean
Species	*Glycine max (L.) Merr.*	Cultivated Soybean

Glycine max is native to Asia, specifically northern and central China and is considered to be one of the oldest cultivated crops. The first evidence for soybean was recorded by Emperor Sheng Nung in 2838 B.C.E. Historical records indicate that the domestication of the crop occurred sometime between 1700 and 1100 B.C.E. [4].

Introduction to the U.S. occurred in 1765 C.E. [5]. Since then, the United States has become the world leader in soybean production, producing 90.6 million metric tons in 2010, with nearly half (43.27 million metric tons) of the yield being exported to other countries. Crop production is measured in bushels. In 2012, 76,104,000 acres were harvested producing 3,014,998,000 bushels of soybeans [6]. The price per bushel in 2012 was $14, making total production value over $43 billion [7]. Soybean is second only to corn in the total area planted in the U.S. as of 2010 [6].

In addition to being a valuable export, soybean has numerous domestic uses. Over 100 uses exist for soybean products for edible consumption and nearly as many, 87, for industrial use. Some edible uses include traditional soy products (such as tofu and soymilk), numerous baked good ingredients, baby food, and livestock feed. On the industrial side, uses are highly varied, ranging from the mundane, such as candles, crayons and cosmetics, to harsher chemicals such as industrial solvents and pesticides. Soybeans provide nearly 70% of dietary proteins and nearly 30% of edible vegetable oils (68% for the United States) in the world [6]. In more recent years, the crop has been recognized as a possible resource for the production of biodiesel, a high-energy alternative fuel [8]. According to the National Biodiesel Board, every bushel of soybean has the potential to produce 48 pounds of protein-rich food and 1.5 gallons of

biodiesel. Since 1999, biodiesel production in the U.S. has been climbing with production reaching a peak in 2008 at 691 million gallons [6].

It was interest in this role as a biodiesel source that prompted a consortium supported by the Department of Energy Joint Genome Institute Community Sequence Program to initiate the soybean genome project that led to the full sequencing of *Glycine max* in 2010. It was thought that knowledge of the soybean genome would provide a means for crop improvements and application towards energy production [8].

Whole genome shotgun sequencing was performed using Sanger protocols and an Arachne algorithm used for assembly of sequence reads. The genome consisted of 1,115 Mega bases, which are assembled into 20 chromosomes. The project resulted in the prediction of 46,430 high-confidence gene loci of which 73% were identified as orthologs of other angiosperms and were assigned to 12,253 gene families. From the angiosperm families, 283 putative legume-specific gene families were identified. Four hundred forty-eight *Glycine max* genes belong to the legume-specific families. In addition, 741 soybean-specific families were identified that could contain potential soybean-specific genes [9]. The whole genome sequencing of soybean has provided scientists with an avenue to understanding the evolution of the species as well as a foundation for further exploration into genomics and proteomics using available bioinformatics tools.

1.2. GENE DUPLICATION AND GENE FAMILIES

Gene duplication is the creation of a duplicate copy of a gene within the genome. Duplication events include retrotransposition, segmental duplication, and whole genome duplication. Retrotransposition occurs when mature RNA undergoes reverse transcription and becomes integrated into the DNA resulting in genes that lack introns. Segmental duplications can result from unequal crossing over or errors in the replication process. Whole genome duplications (WGD) are rare events, but provide opportunity for great changes to be made in organismal complexity and behavior [10]. A WGD results in a two-fold expansion of the genome and a polypoidy state. It is suggested that this state contributes to increased adaptability and tolerance in extreme conditions. Genome sequencing of numerous plant species has revealed synteny between species, or the retention of tracts of homologous genes between species of the same family. Based on this information, the timing of major duplication events can be predicted. A WGD event is estimated to have occurred in a number of plant species shortly after the mass extinction event at the end of the Cretaceous period, around 65 million years ago (Mya), providing those plants with an evolutionary advantage in a time when selective pressures were leading to the extinction of other species [11].

Glycine max (*Gm*), *Medicago truncatula* (*Mt*), and *Lotus japonicum* (*Lj*) are three species of the Papilionoid subfamily of legumes that have completed genome sequences. This subfamily diverged from the two other legume subfamilies around 60 Mya. By analyzing blocks of synteny between species on a dot-plot matrix, scientists have concluded that a WGD occurred around 58 Mya, prior to the divergence of *Mt* and *Lj* from one another [11]. Speciation creates orthologous genes, ones that share

homology and were created by the splitting of a lineage, between the species [12]. A comparison of *Gm* and *Mt* on a similar dot-plot matrix revealed synteny between the species but with pairs of syntenic blocks in *Gm* corresponding to single blocks in *Mt*. This indicates a WGD in *Gm* after its divergence from *Mt* [11]. A gene duplication event in a single genome produces paralogous genes [12].

Gene duplication events can produce closely related genes in a genome that encode the same, or very similar, products (generally proteins). Two or more genes that meet this criterion are considered to be a gene family. Members of a gene family can be located on the same chromosome or dispersed throughout the genome [13].

Most genes in the genome are vulnerable to selective pressures and because of this, most nucleotide changes are deleterious. The creation of a copy of a gene provides an opportunity for a new gene, one with a novel function, to arise. As long as one copy of a gene maintains the original function, the other copy can accumulate mutations without negatively affecting the fitness of the organism. The development of a novel function by one copy of a gene while the other retains the original function is known as neofunctionalization. In some instances, the original function is not wholly retained in either copy; rather, it is divided between the copies so that each copy contributes a portion of the function. This is referred to as subfunctionalization [10].

Duplicated genes have several possible fates. Those that decrease fitness are eventually lost. Duplications that do not provide any benefit or harm will be subject to drift and eventually fixed or lost. A duplicated gene that provides an advantage against selective pressures will eventually be fixed. Duplicated genes that are lost can

become non-functional, referred to as a pseudogene, through the accumulation of degenerative mutations, or be physically lost from the genome. Fixation is not a common event; only 1 in every 100 genes becomes fixed every million years. In fact, studies show that the window of opportunity for a newly duplicated gene to become fixed in a population before degradation commences is very small on the evolutionary scale, somewhere in the order of 4 million years [10].

Mutations become introduced into and accumulate in the DNA, altering the percentage of a gene that retains identity with its paralog(s). One type of mutation is a base substitution which can come in two forms: synonymous and nonsynonymous. Synonymous substitutions are "silent," no change occurs in the amino acid that is encoded by the codon. Nonsynonymous substitutions arise from the replacement of a nucleotide that results in the DNA encoding a different amino acid or the creation of a transcription termination signal [13]. Synonymous substitutions accumulate at a fairly steady rate; therefore, the calculation of the substitutions per site (Ks) between paralogs can be used as a measure of time since a duplication event [14]. In *Gm*, 31,264 of the 46,430 high-confidence genes exist as paralogs with a Ks value of approximately 0.13 synonymous substitutions per site. This corresponds to the *Gm*-specific WGD at 13 Mya according to the 2010 Schmutz et al. Nature article *Genome sequence of the palaeopolyploid soybean* [9]. Schmutz et al. analyzed synonymous substitution rates between soybean gene families with no more than six members that have a paralog on another chromosome and found two peaks. Peaks appear on the graph at the values 0.13 and 0.59 Ks. The peaks indicate a high percentage of homologous pairs of genes within the *Glycine max* genome. The Ks values correspond

to two separate duplication events, the 13 Mya and 59 Mya WGDs. Each value does not correspond directly to a different point in time; rather, a range of values can be correlated to a particular event. The range of Ks values surrounding the 0.13 peak correspond to the 13 Mya WGD. The second peak, and flanking values, in the comparison of *Glycine max* paralogs appears at Ks value 0.59 which corresponds to the Papilionoid, or legume, genome duplication at 59 Mya [8]. This evidence correlates well with the synteny dot-plot matrix analysis.

Analysis of paralogs in *Glycine max* is further complicated by the fact that soybean is an allotetraploid. Autopolyploidy can result from somatic doubling in a single species, whereas allopolyploidy results from hybridization of different but related species followed by somatic doubling to create a fertile derivative [15]. At some point after the Papilionoid divergence, two species of legumes existed (proto-*G. max* species) that had n = 10 chromosomes each. Enough molecular compatibility existed for the pollen from one species to fertilize the egg of the second, resulting in a hybrid plant. Each species would contribute 10 chromosomes, however the hybrid would be sterile due to the absence of complete homologous chromosomal pairs. Upon somatic doubling, each of the 10 + 10 chromosomes becomes duplicated, chromatids following S phase become homologs after a failed anaphase, producing 20 pairs of chromosomes and reinstating the ability of the organism (*Glycine max*) to reproduce. This process is outlined in Figure 1.1 below.

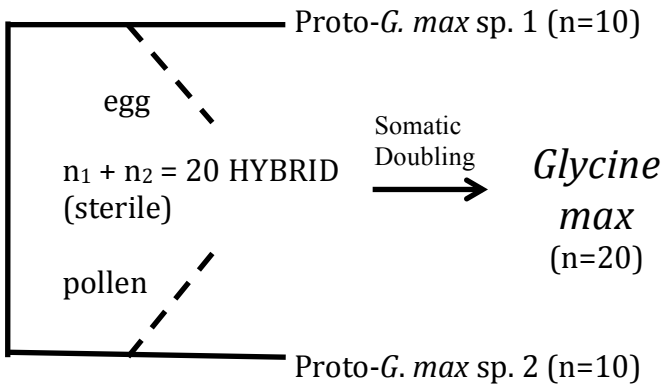


Figure 1.1. The hypothesized allotetraploid event that produced *Glycine max*.

1.3. EVIDENCE OF GENE EXPRESSION

Predicting the presence of a gene does not prove that the encoded sequence becomes a functional protein or that it is even transcribed.

1.3.1. ESTs. An expressed sequence tag (EST) is a short/partial sequence derived from cDNA [16]. A cDNA is created from the reverse transcription of mature mRNA, which is a record of protein-coding DNA. ESTs are created by sequencing a short segment (usually 200-500 bp) of a cDNA from the 5' or 3' end [17]. Since ESTs are created from an experimentally obtained product of gene expression, they provide empirical evidence of transcription at minimum, and possibly protein synthesis. High levels of synthesis of a protein require high levels of transcription of the encoding gene, i.e. numerous mRNA transcripts. Since ESTs provide a record of transcript production in a particular cell, tissue, or organ of an organism [17], they are a reflection of gene expression and may provide additional evidence for differential expression patterns if tissue treatment is known.

Because they are records of a transcript, the EST sequences represent post-spliced sequences; the introns are removed. This presents a challenge for determining the encoding gene. Despite this, the creation of ESTs is an inexpensive means of finding new genes, gaining information about expression, and creating genome maps [17], making it a convenient tool for studying the genome of organisms that have not yet been sequenced [18].

EST records are produced and maintained by Genbank and are available through the EST database at NCBI (dbEST) [19].

1.3.2. Consensus Data. It has been discovered that some DNA elements and signal sequences have motifs that are conserved across many taxonomic boundaries. Collections of such consensus sequences can aid in the identification of possible regulation of novel genes and can provide additional evidence regarding gene expression.

1.3.2.1. Promoter elements. The promoter is a region of sequence located upstream of the transcription start site. Elements within the promoter contribute to the inherent affinity of the region for RNA polymerase, an enzyme that relies on promoter recognition by necessary accessory proteins (transcription factors) to initiate transcription in the correct location [13].

The most common element, the TATA box, can be found in up to 50% of eukaryotic genes and is generally located between 23 and 33 base pairs (bp) upstream from the transcription start site. The eukaryotic TATA box consensus sequence is TATAAA, however there are numerous variations of this sequence found between species. An Inr box is another element common to eukaryotic genes (40-65%) and

straddles the transcription start site. This element is C/T rich and may serve as an important protein binding site in the absence of a TATA box. The downstream promoter element (DPE) is located 23 and 33 bp after the transcription start site. The CAAT box, with the consensus sequence GGNCAATCT, is located between 40 and 100 bp upstream from the transcription start site if present [16]. The significance of the presence of these elements is based on the need for binding sites for proteins that assist in RNA polymerase positioning. If transcription factors and other accessory proteins that require these elements cannot bind because the elements are deleted or mutated, transcription cannot initiate, and a non-functional gene (pseudogene) is the result. Thus, the presence or absence of promoter elements can provide evidence for the expression of a predicted protein-coding sequence.

A number of web-based promoter element prediction tools exist. For the scope of this project, the Plant Cis-Acting Regulatory DNA Elements (PLACE) database was the preferred search tool. FgenesH, a gene prediction tool that is used to construct gene models, also predicts a promoter region but not specific elements.

1.3.2.2. Polyadenylation signals. The polyadenylation (poly-A) site is a series of sequences that signals an endonuclease to cleave a transcript at a specific site located 10 – 30 nucleotides downstream from the polyadenylation signal sequence [13]. Eukaryotic polyadenylation sites lie in the 3' UTR of the genes that encode the transcripts and are composed of three cis-elements—the poly-A signal, the cleavage site, and a downstream element (DSE). No consensus data has been found for the DSE; it is known only to be a U/GU-rich region located 10 – 15 nucleotides downstream from the cleavage site [20]. Once an mRNA molecule is cleaved it can

undergo essential processing, which includes the addition of a tail composed of repeating adenosine residues. This is a vital eukaryotic process that stabilizes the molecule and in doing so, promotes efficient translation of the mature transcript into a protein [13].

Studies have revealed that the poly-A signal is a highly conserved hexanucleotide in animals with the consensus sequence AWUAAA, where W stands for U or A [20]. However, conserved motifs of polyadenylation signals in plant species are less conserved and far more difficult to identify. For example, in a 1987 study, the sequence AAUAAA was only found to exist in 10% of the transcripts produced from Arabidopsis genes [21]. In addition, it has been observed that plant genes may contain multiple poly-A sites [22] and studies of Arabidopsis reveal that the DSE may not be present in plants [20].

In 2013, Sherstnev et al. revealed that *Arabidopsis* does contain multiple poly-A sites, but there are preferred profiles associated with cleavage sites. For a small quantity of cleavage sites, the most common motif was in fact the consensus sequence AAUAAA, with the preferred location of 19 nucleotides upstream from the cleavage site. For others, the poly-A signal was a similar hexamer, differing in the position of a single residue. In addition, a U-rich sequence was found to consistently reside 7 nucleotides upstream from the cleavage site (USE), as well as a short A-rich sequence and a U-rich DSE. It was concluded that the presence and multi-functional purpose of the U- and A-rich regions might account for the decreased use of a consensus sequence at the poly-A signal [23]. Despite growing knowledge of the poly-A site of plant species, currently the best resource for the identification of plant poly-A sites in

plants is the analysis of ESTs that contain polyadenylation tracts [22]. However, specific algorithms have been, and are being, developed to predict these sites and include a Generalized Hidden Markov Model (GHMM), Adaboost, length-variable second order Markov model (LVMM2) [22], and Generalized Hidden Markov Model-Real Wavelength Transform (GHMM-RWT) [24].

1.3.2.3. Intron/exon borders. Research into the splicing mechanism led to the discovery of consensus sequences at intron/exon borders that are recognized by snRNAs within the splicing machinery. At the 5' splice junction of the primary transcript, the consensus sequence is CAG/GUAAGU (/ indicates the border between exon and intron). At the 3' splice junction the consensus sequence is UUUUCCCUCCAG/GU. The encoding DNA contains the 5' and 3' nucleotide dimers GT and AG respectively. Figure 1.2 shows the sequence logo as generated by a motif analysis program for retained introns, constitutive exons, and skipped exons [16].

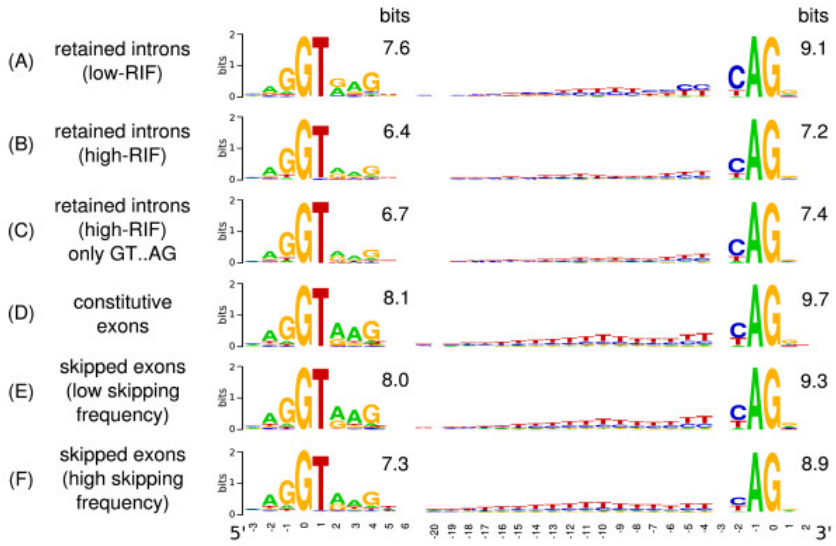


Figure 1.2. Intron/exon border motifs [16]. Original source of illustration: Figure 1. From Sakabe NJ, de Souza SJ. Sequence features responsible for intron retention in human. BMC Genomics 2007;8.

The intron border consensus data is ubiquitous among eukaryotic organisms and is summarized in Figure 1.3.

Exon 1	5' s.j	Intron	3' s.j.	Exon 2
G	GT		AG	G

Figure 1.3. Eukaryotic intron border consensus sequences [16].

1.3.2.4. Splicing signals within introns. In animal species, there are four signals intrinsic to the intron that mediate splicing activity. Two of these sequences are the 5' and 3' splice junction consensus dimers, GT and AG respectively. There are two additional signals found near the 3' splice junction—the polypyrimidine tract and a branch point located 17-40 nucleotides upstream of the 3' end of the intron [25]. Spliceosome components recognize and bind these signals. The 5' splice junction sequence is recognized by the U1 snRNP, the 3' splice junction sequence is recognized by the $U2AF^{35}$ protein, the polypyrimidine tract is recognized by the $U2AF^{65}$ protein, and the branch point is recognized by the U2 snRNP [26]. $U2AF^{35}$ and $U2AF^{65}$ are U2 auxillary factor protein subunits of 35 kDa and 65 kDa of the heterodimer U2AF. $U2AF^{65}$ directly binds the polypyrimidine tract as well as another protein, SF-1/mBBP. The complex of these proteins promotes binding of the U2snRNP unit to the branch point sequence [28].

Comparative studies between animals and angiosperms revealed some key differences in gene structure. Plants have shorter genes with fewer numbers of exons and shorter introns [25]. In addition, the branch point [25] and polypyrimidine tract [28] are not always identifiable due to the 3' region of the intron being rich in U/A

residues [25/26]. U/A composition is essential to the functionality of an intron for splicing. It was reported by Goodall and Filipowicz in 1991 that the minimum UA content for introns in dicots is 59% for splicing efficiency [29]. The U richness of intronic elements allows for dual functionality. In the presence of a branch point, the U-rich element can serve as a polypyrimidine tract and in the absence of a branchpoint, it can serve as a UA-rich element [28].

Studies have been conducted to determine the optimum consensus sequence for the branch point. The loosely conserved consensus sequences for plant and vertebrate branch points (of the pre-mRNA) were proposed in 1986 to be CURAY and YURAY (where Y stands for C or T and R stands for A or G [32]) [30] and were later confirmed in a 2002 mutational analysis [31].

Due to the fact that intron retention is the most common form of alternative splicing event in plants, scientists propose that the signals for splice site recognition are likely located in the intron. In addition to the conserved sequences described previously, regulatory elements can also mediate splicing activity. These elements are referred to as splicing regulatory elements (SRE) and can exist in the intron or exon as enhancers or silencers, promoting or preventing the use of a splice site. There has been very little effort to date to uncover SRE's in plants [33]. Some computational tools have identified putative exonic splicing enhancers (ESE) in *Arabidopsis thaliana* [34] such as a GAAG repeating region of the intron known to bind the regulatory protein SCL33 [33]. ESEs are the most commonly studied SRE, however, their study is limited primarily to mammalian species [26].

1.3.3. MicroRNA. MicroRNA (miRNA) is a short (~22 nucleotides long) type of non-coding RNA (ncRNA) that is typically involved in post-transcriptional regulation of gene expression through mRNA degradation or translational repression. Micro RNA can typically be found in a symmetrical structural formation, such as a hairpin or cloverleaf, that is the result of dyad symmetry (inverted repeats) within the coding sequence. Inverted repeats are thought to be the product of inverted DNA from a duplication event. If the function of a miRNA is important, sequence and structure are conserved [35].

MicroRNA-encoding genes are present on all 20 soybean chromosomes and are predominantly intergenic [36]. MicroRNAs can be located in the introns of functional genes and these intragenic miRNAs have been observed being expressed through parent gene preferential expression in root tissue. This is likely due to the strong role that legume-specific and nodulation-regulated miRNAs play in root nodule development [37]. Legume- and soybean-specific miRNA families tend to be smaller and contain fewer members, whereas highly conserved miRNA families tend to be larger with more members [36]. In soybean, most of the legume-specific miRNA families produce a 21 nucleotide mature miRNA molecule. Also, soybean miRNAs exhibit a preference for U at the 5’ most nucleotide of the mature molecule [36].

Computational analytical tools such as miRseeker, miRScan, miRRim [35], and FOMmir [38] exist that use trained algorithms to predict miRNA sequences [35]. In 2010 Kandoth et al. presented a novel approach to identify miRNA precursors that utilizes an algorithm to search for inverted repeats and then filters the results using criteria such as density and length of symmetrical area, and GC content [35].

1.3.4. Dyad Symmetry. Dyad symmetry exists in a double-stranded DNA sequence when a segment of sequence from one strand can be rotated 180 degrees and match the sequence of the complimentary strand of the same segment. This implies intramolecular base pairing capability, which, for example, can result in the formation of structures such as hair-pin loops. In order to determine whether sequences exhibit dyad symmetry, a test was developed to determine what type of output dyad symmetry would produce in a dot plot matrix. A random selection of nucleic acids was assembled into a 50 nt long sequence with a 12 nt internal segment that exhibited dyad symmetry. This sequence was plotted against itself, its compliment, and its reverse compliment in a dot plot matrix. Plotting against the reverse compliment resulted in a distinct graph that would indicate dyad symmetry. The reverse compliment sequence can be obtained using a web-based translation tool.

1.4. DATABASES AND OTHER BIOINFORMATIC TOOLS

1.4.1. Proteins: PFAM ,Panther, KOG, and PDB. PFAM is a protein family database in which a family being defined as "sets of protein regions that share a significant degree of sequence similarity" [39]. PFAM contains both manually curated and automatically generated families produced from hidden Markov model profiles created and searched against the UniProtKB database. The ultimate goal of the PFAM database is to assemble a set of annotated families that can be used for genome-annotation and protein studies [39].

PANTHER (Protein Analysis Through Relationships) is another gene family database that classifies proteins according to family/subfamily, molecular function,

biological process, and pathway [40]. It provides three types of annotation—subfamily membership, protein class, and gene function. The annotations are linked to nodes on a phylogenetic tree for each family. PANTHER was originally developed in anticipation of the first sequencing of the human genome. The website provides tools for the functional analysis of genes and proteins [41]. PANTHER also annotates according to gene ontology [41] and is involved with the Gene Ontology Reference Genome Project [40].

KOG (euKaryotic Orthologous Groups), which is an update of the original system COG (Clusters of Orthologous Groups) [42], is a collection of proteins from eukaryotic genomes that are classified according to four functional groups and clustered according to orthology and parology [43].

The world-wide Protein Data Bank (PDB) is a publically-accessible database of macromolecular structural data supported by a collaboration of numerous international research organizations including the Research Collaboratory for Structural Bioinformatics (RCSB) which is managed by Rutgers, the State University of New Jersey, and the San Diego Supercomputer Center at UCSD; the Macromolecular Structural Database (MSD) at the European Bioinformatics Institute (EBI); and the Protein Data Bank Japan (PDBj) [44].

1.4.2. Phytozome. Phytozome is a hub for the comparative studies of plant families and evolution. Phytozome is supported by the Department of Energy's Joint Genome Institute (JGI) and the Center for Integrative Genomics (CIG). Currently running in version 9.1, it provides users access to 41 sequenced green plant genomes. Annotations based on PFAM, KOG, KEGG, and PANTHER assignments, as well as

publicly available annotations from RefSeq, UniProt, TAIR, and JGI, are provided where available. Phytozome regularly updates genomic information as it becomes available. The web portal is user friendly and offers tools such as a genome browser and basic local alignment tools. By selecting a specific gene or transcript, the user gains access to basic information, sequence data (genomic, transcript, coding, and peptide sequences), protein homologs and gene ancestry [45].

1.4.3. NCBI. The National Center for Biotechnology Information (NCBI) was established in 1988 as a division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH) due to recognition of the need for computerized information processing methods in biomedical research. The organization's mission became "finding new approaches to deal with the volume and complexity of data and in providing researchers with better access to analysis and computing tools to advance understanding of our genetic legacy and its role in health and disease [46]." The specific goal of NCBI was focused on:

> creating automated systems for storing and analyzing knowledge about molecular biology, biochemistry, and genetics; facilitating the use of such databases and software by the research and medical community; coordinating efforts to gather biotechnology information both nationally and internationally; and performing research into advanced methods of computer-based information processing for analyzing the structure and function of biologically important molecules [46].

NCBI has a multi-disciplinary research group that supports the GenBank DNA sequence database and a number of other databases including the database of expressed sequence tags (dbEST), Online Mendelian Inheritance in Man (OMIM), the Molecular Modeling Database (MMDB) of 3D protein structures, the Unique Human Gene Sequence Collection (UniGene), a Gene Map of the Human Genome, the

Taxonomy Browser, and the Cancer Genome Anatomy Project (CGAP), in collaboration with the National Cancer Institute [47].” NCBI shares data with the European Molecular biology Laboratory (EMBL) and the DNA Database of Japan (DDBJ). NCBI is a central hub for access to numerous automated gene and protein analysis tools such as BLAST, RefSeq, 1000 genomes browser, an open-reading frame finder (ORF-finder), and numerous taxonomy and classification tools. Entrez, a data-mining tool offered by NCBI, provides users with access to sequence, mapping, taxonomy, and structural data [47].

1.4.4. DNA Subway. DNA Subway is supported by the Dolan DNA Learning Center as part of iPlant Collaborative. It provides a bioinformatics workspace for users to analyze sequence data, create annotations, and perform phylogenetic analysis [48]. The user-friendly interface facilitates the use of analytical tools by making user options available in a step-by-step manner through a mapped out pathway. Users can choose pathways for gene annotation, genome prospecting, sequence relationship determination, and next generation sequencing. Within the gene annotation tract, the user can use predictive algorithm programs (Augustus and FgenesH) to predict genes from sequence inputs, upload nucleic acid and peptide sequences for comparison, and construct gene models in APOLLO, a curation and annotation tool. Augustus predicts a 5’untranslated region for possible promoter location and FgenesH predicts the most likely start codon. This makes both tools uniquely valuable as predictive tools [49].

1.4.5. Phylogeny.fr. Phylogeny.fr is a free website developed for use by non-specialists for the construction and analysis of phylogenetic relationships between sequences. It provides a workflow for users that can be set up to run without any

decision-making on behalf of the user or with full control of program selection and parameter settings by the user. The output of the workflow is the result of sequence processing by a multiple alignment, a curation, a phylogeny, and a tree-rendering tool [50].

1.4.6. ExPASy. The Expert Protein Analysis System (ExPASy), which is supported by the Swiss Institute of Bioinformatics (SIB), is a web portal that offers access to databases and other tools for the purpose of analysis within the areas of genomics, proteomics, phylogeny, systems biology, population biology, etc. ExPASy is the main host for databases developed by the SIB including PROSITE, a database of protein families and domains that aids in the identification of novel sequences through known protein family annotations [51]. ExPASy databases are cross-referenced with other biological resources all over the world and are updated frequently [52].

1.4.7. MEME. MEME (Multiple Ems for Motif Elicitation) is a member of the MEME suite, a web server that is funded by the National Biomedical Computation Resource and hosts tools for motif-based sequence analysis [53]. The MEME tool employs an algorithm to identify motifs in a set of DNA or protein sequences. The technique involves "expectation maximization to fit a two-component finite mixture model" to a set of unaligned sequences. The only required parameter is the selection of motif width [54]. This tool is only capable of generating motifs from input sequences and does not compare them to known motifs.

1.4.8. CLUSTALw. CLUSTALw is a web-based multiple sequence alignment tool that is supported by GenomeNet, a network of database and computational

services operated by the Kyoto University Bioinformatics Center [55]. This tool aligns multiple nucleic acid or peptide sequences for phylogenetic analysis [56].

1.4.9. PAL2NAL. PAL2NAL is a web-based (also downloadable), program that converts a peptide multiple sequence alignment to a codon alignment by referencing the peptide sequences to their coding DNA sequences. This is accomplished by reverse translation of the peptide sequence to a DNA sequence based on regular expression patterns and subsequent comparison of the reverse translation product to the input DNA coding sequence to find corresponding coding regions. The PAL2NAL server accepts the multiple alignment data in FASTA or CLUSTAL format. The resultant codon alignment is useful for calculating synonymous and non-synonymous substitution rates [57].

1.4.10. SNAP. The Synonymous Non-synonymous Analysis Program (SNAP) v1.1.1 is a tool of the HIV sequence database that calculates synonymous and non-synonymous substitution rates from a codon alignment [58/59].

1.4.11. PLACE. PLACE (Plant Cis-acting Regulatory DNA Elements) is a database of motifs common to known cis-acting regulatory DNA elements of vascular plants. Motifs are found by uploading a nucleic acid sequence into a signal scan search. All motifs that match or are similar to known motifs will be listed with the promoter element/site name, location of the starting nucleotide, whether it is located on the + or – strand, the motif/signal sequence, and a PLACE identification number (PLACE identifier). The PLACE identifier provides information about the source and expression. Motif descriptions are available with accession numbers for any information cross-referenced with PubMed or GenBank/EMBL/DDBJ [60/61].

1.4.12. CELLO. CELLO is a subcellular localization prediction tool that uses a support vector machine (SVM) classification system to predict the location of cellular activity of a peptide sequence based on amino acid composition, di-peptide composition, partitioned amino acid composition, and physio-chemical properties of amino acids. Compared to other subcellular localization prediction methods, CELLO displays the highest prediction accuracy [62/63/64]. Subcellular localization can be useful for inferring protein function since function is usually related to the location of employment [64].

1.4.13. PSIPRED Protein Sequence Analysis Workbench. PSIPRED (Position-Specific Iterated Prediction) is a secondary structure prediction method that annotates sequences with the location of key structural features such as coiled coil, helical, and sheet domains. It employs a two-stage neural network based on position-specific scoring matrices produced by PSI-BLAST [65]. It is supported by the UCL Department Of Computer Science Bioinformatics Group and is currently running version 3.3 [66].

PSIPRED and numerous other recognition programs can be accessed and employed simultaneously through the PSIPRED Protein Sequence Analysis Workbench at http://bioinf.cs.ucl.ac.uk/psipred/. Two of these tools, fold recognition and fold domain recognition, are useful when analyzing a novel peptide sequence. GenTHREADER is a rapid fold recognition tool that can be used to infer tertiary structure. It employs a simple neural network to combine sequence alignment score, length information and energy potentials and threads them into a single score. The

score represents the relationship between two proteins according to CATH (a protein structure classification database) designation [67].

CATH is a hierarchical domain classification system. The four levels of hierarchy include [68]:

- Class - classified according to secondary structure.
- Architecture – classified according to orientation of secondary structure in 3-dimensional space.
- Topology – classified according to fold groups and connection between secondary structure.
- Homologous superfamily – classified according to domain group that is thought to share a common ancestor.

pGenTHREADER (Parametric-GenTHREADER) and pDomTHREADER (parametric-DomTHREADER) are improved versions of the GenTHREADER program used for protein sequence alignment and recognition. Both use the same core algorithm. The fold recognition algorithm is guided by 20 parameters: four for profile-profile scoring, nine for secondary structure scoring, six for gap penalties, and a weighted score. Both versions use the profile-profile score to produce a measure of confidence score [69]. pGenTHREADER uses profile-profile alignments and secondary structure prediction as input to improve accuracy. pDomTHREADER produces domain alignments in an effort to improve accuracy of superfamily discrimination [70].

1.4.14. DNA Dot Plots. Dot plots provide a simple, graphical method of analysis of sequence similarity between two different sequences, repeats within a

single molecule, and potential for intra-molecular base pairing (tertiary structure). The comparison matrix is employed by the movement of a window (number of residues being considered at one time; defined by user) along two input sequences. The user defines the mismatch limit, which is the number of mismatched residues that can exist within a window in order to still be defined as a match [71]. For example, if the window size is 5 and the mismatch limit is 1, a single mismatch can exist in a 5 base sequence and it is still considered a match. Under these parameters, the existence of 2 mismatches in a 5 base sequence is not considered a match. If the matrix determines a match at a single point in the comparison a dot is placed on the matrix. The appearance of diagonal lines is an indication of sequence similarity. The appearance of parallel lines is often a sign of repeating sequences.

1.4.15. I-TASSER. The I-TASSER server is an automated online platform for predicting the structure and function of proteins. Using an amino acid sequence as the query, the program performs a four-stage protocol to determine structure. The first stage involves threading the query sequence against solved structure databases to find template proteins with similar structure or motifs. A multiple alignment of structural homologs is used to create a sequence profile, which is used to predict secondary structure and will assist in threading the query against the PDB library. The top templates are ranked according to specific score criteria and are further scrutinized using tests of statistical significance. The second stage involves threading alignments of fragments from templates in order to assemble regions that align well into structural conformations. The third stage is model selection and refinement. One function of this stage is to refine global topology through removal of steric clashes in the model. In the

fourth stage involves functional annotation prediction through a comparison of the predicted model with proteins of known structure and function in PDB [72].

Each stage produces multiple categories of scores that drive the process—confidence scores, identity scores, statistical scores, etc. The primary scores for assessing the predicted 3D model are the C-score, or confidence score, and the TM-score, a measure of the quality of the final model. Selection criteria for correct model topology are a TM-score greater than 0.5 and a C-score greater than -1.5, because the false-positive and false-negative rates are low [73].

I-TASSER has been ranked the number one server for structural and functional prediction of proteins in many recent community-wide Critical Assessment of Structure Prediction (CASP) experiments [73]. The server ranked number one in CASP7, CASP8, CASP9, and CASP10 for structural prediction and in CASP9 for functional prediction as well [74].

2. MATERIALS AND METHODS

2.1. BLAST

Sequences were aligned against a target database using a Basic Local Alignment Search Tool (BLAST). The types of BLAST searches that were conducted are outlined in Table 2.1. Query sequences are input in FASTA format (>glyma03g . . .). For the tblastn searches (predominant BLAST type used), algorithm parameters were set as follows: Max target sequences, 100; Expected threshold, 10; Scoring matrix, BLOSUM62; Gap costs, Existence: 11 and Extension: 1. For the blastn searches, algorithm parameters were set as follows: Max target sequences, 100; Expected threshold, 10; Word size, 28; Match/mismatch scores, 1,-2; Gap costs, Linear. For the scope of this research effort Phytozome was the preferred genome browser, which has a built in BLAST application, but the BLAST tool at the National Center for Biotechnology Information (NCBI) was also utilized.

Table 2.1. Summary of database and query requirements for utilized BLAST programs.

BLAST Program	Query Sequence	Database Searched
blastn	Nucleic acid sequence	Nucleotide
tblastn	Conceptually translated peptide sequence	Protein

2.2. SEQUENCE ALIGNMENTS

Sequence alignments were conducted using CLUSTALW. Both DNA and protein pairwise alignments and multiple alignments were performed using the

slow/accurate alignment option. The protein alignment inputs consisted of the conceptually translated peptide sequences of all genes to be aligned in FASTA format. The nucleotide alignment inputs consisted of the nucleic acid sequences of all genes to be aligned in FASTA format. CLUSTAL was chosen as the output format. Parameter settings are outlined in Table 2.2.

Table 2.2. CLUSTALW parameters.

	Pairwise Alignment		Multiple Alignment	
	DNA	Protein	DNA	Protein
Gap Open Penalty	15	10	15	10
Gap Ext. Penalty	6.66	0.1	6.66	0.05
Weight Matrix	IUB	BLOSUM	IUB	BLOSUM

2.3. CHOICE OF GENE FAMILY AND IDENTIFICATION OF MEMBERS

A table of candidate gene families was compiled by cross-referencing information from the PFAM database and Supplementary Table 5 [9]. PFAM was filtered for all gene families in *Glycine max* annotated as unknown function and listed in descending order according to the gene count. Gene families containing 10 or fewer members were further screened as follows. Supplementary Table 5 in Schmutz et al. [9] provided additional information concerning the quantitative relationship between 10 plant species, including *Glycine max*, and a gene address for a single gene in *Glycine max* for each family. The number of putative genes in a family is compared across ten species including *Vitis vinifera* (Vvi. common grape), *Populus trichocarpa* (Ptr, cottonwood or poplar tree), *Medicago truncatula* (Mtr, barrel medic), *Glycine max* (Gma, soybean), *Arabidopsis thaliana* (Ath, thale cress), *Arabidopsis lyrata* (Aly,

rock cress), *Carica papaya* (Cpa, papaya or pawpaw), *Sorghum bicolor* (Sbi, sorghum), *Zea mays* (Zma, corn), *Brachipodium distachyon* (Bdi, purple false brome), and *Oryza sativa* (Osa, rice), and is displayed in a ratio as Vvi:Ptr:Mtr:Gma:Ath:Aly:Cpa:Sbi:Zma:Bdi:Osa. Focus was placed on those families that had a PFAM functional annotation of UNKNOWN, contained ten or fewer members in *Glycine max*, and displayed expansion in *Glycine max* relative to the other species in the gene family ratio, i.e., fewer than 3 members in any other species.

A conceptually translated peptide sequence, obtained from Phytozome, from a single family member of each potential family was used as an initial query in a BLAST. The query was compared to all sequences in the *Glycine max* genome. Those hits that produce a potential transcript were designated “model genes” and those sequences that do not correspond to a putative gene (algorithm-predicted) were designated “non-models” (referred to as non-coding sequences throughout this manuscript). All hits producing a possible transcript were subsequently used as a query using the same BLAST parameters. A family was arbitrarily defined as a group of model genes in which each gene registered as a hit for all other potential family members.

Based on data collected from all aforementioned filters, three families were of particular interest, all exhibiting high similarity between members. The sequences of those particular family members were placed into MEME (Multiple Em for Motif Elicitation), a motif-based sequence analysis tool. The input was conceptually translated peptide sequences of all genes in a family in FASTA format. MEME parameters were set to find between 2 (minimum) and 20 (maximum) sites per

sequence. The width of each motif was limited to a minimum of 6 and a maximum of 50 with a maximum of 3 motifs per sequence. From those motifs, any conserved sequences of ten or more consecutive amino acids were used as queries in a BLAST search of the *Glycine max* genome to ensure that all members of each family had been discovered.

The *Glycine max* gene family containing a putative gene at the location 08g39410, displayed 3 strongly conserved motifs, 2 of which exceeded fifty amino acids in length and four out of five family members contain all three motifs.

From this point forward all research was limited to the gene family that contains glyma08g39410 and the members of the family hereafter referred to as LJFgene3, LJFgene14, LJFgene1, LJFgene8, LJFgene9, LJFnm19, LJFnm11, and LJFnm12.

2.4. EVOLUTIONARY AND EXPRESSION ANALYSIS FOR GENE MODEL CONSTRUCTION

2.4.1. Neighbor Gene Analysis. The predicted genes adjacent to each family member that lie within 50 kbp on either side of it on the chromosome were examined to identify any synteny that might exist between family members. The following information was recorded for each gene adjacent to a gene family member:

- Gene address.
- 5’ or 3’ placement on strand relative to the query (with consideration given to the orientation of the query on the + or – strand.)
- Distance from the query gene in kbp.
- Annotated function, if any.

The non-coding sequences (LJFnm19, LJFnm11, and LJFnm12) that resulted from the original BLAST searches were also analyzed using this method.

2.4.2. EST's. The NCBI website contains the information for all expressed sequence tags (EST's) for *Glycine max*. The peptide sequences of all gene family members were individually aligned through a BLAST search against known *Glycine max* EST's to determine which EST's originate from this family. The conceptually translated peptide sequences of each gene were used as query against the *Glycine max* EST database. The query sequence used was in FASTA format. The database was specified as expressed sequence tag (est) and the organism specified as *Glycine max* (taxid: 3847). The max target algorithm parameter was adjusted to 1000 to ensure all EST's were found.

In order to verify that each EST does represent the gene in question, the nucleic acid sequence of each of the resultant EST sequence was used as a BLAST query back to the *Glycine max* genome. The nucleic acid sequence of the EST was obtained from NCBI through searching the corresponding accession number.

The top result is the strongest score and the gene model/sequence that the EST corresponds to. If the top hit corresponds to the gene that was the original query, then that EST belongs to that gene. If the top hit is not the original query, the EST belongs to another gene. All EST accession numbers and max scores corresponding to gene family members were recorded.

A library was created for EST's belonging to gene family members. Information was gathered regarding the accession number, cultivar, tissue, and treatment.

2.4.3. Synonymous and Nonsynonymous Substitution Rates. SNAP (Synonymous Non-synonymous Analysis Program) v 1.1.0, a web-based tool, was used to determine the synonymous and non-synonymous substitution rates between all genes.

The first step was creating a multiple alignment of the putative peptide sequences of the genes in the family. The multiple sequence alignment was then used as input file 1 for creating a codon alignment using PAL2NAL (protein alignment to nucleic acid alignment). Input file 2 for this tool was the coding nucleotide sequences of all gene family members in FASTA format. The codon table was set as universal code, gaps and mismatches were not removed, and the output format was set as CLUSTAL.

The codon alignment output was transferred to the SNAP v 1.1.0 program of the HIV sequence database at www.hiv.lanl.gov. SNAP generated a results table with synonymous and non-synonymous substitution rates for pairwise comparisons of every gene in the family.

2.4.4. *Glycine max* Family Phylogeny. The following tools were utilized to perform the specific tasks for phylogenetic analysis: ClustalW for the multiple alignment, Gblocks for alignment curation, PhyML for phylogenetic tree construction, and Drawtree for phylogenetic tree visualization. The conceptually translated peptide sequences for all genes in the family in FASTA format were provided as input.

2.4.5. Plant Species Family Phylogeny. A cross-species comparison was conducted for each gene in the family by conducting a BLAST search using the conceptually translated peptide sequence of each gene against 9 plant genomes with

varying phylogenetic relationships to *Glycine max*. The species chosen for the comparison were *Medicago truncatula*, *Phaseolus vulgaris*, *Arabidopsis thaliana*, *Zea mays*, *Oryza sativa*, *Sellaginella moellendorfii*, *Physcomitre patens*, *Chlamydomonas reinhardtii*, and *Volvox carteri*. The resultant transcripts and their percentage similarity were recorded. The peptide sequences from these hits were included with the *Glycine max* sequences to estimate phylogenetic trees. Two trees were generated, one with the five *Glycine max* genes and the transcripts from *Medicago truncatula* and *Phaseolus vulgaris* (closely related legumes) and a second with the five *Glycine max* genes and all of the aforementioned species' transcripts.

2.4.6. Constructing Gene Models. The specific tools utilized for gene model construction were as follows.

2.4.6.1. Predicting gene models. For each gene family member, a 10 kbp segment of the chromosome with the gene model at near center in FASTA format were analyzed. Repeat masker was used to eliminate/block repetive sequences within the query that could slow the analysis. Augustus and FgenesH, two predictive programs that use unique algorithms, were employed to predict the presence of a gene within the input sequence.

The nucleic acid sequences for all the ESTs for each gene family member were aligned against the 10kbp segment. Aligned ESTs and predicted models were viewed using APOLLO, a model building application.

2.4.6.2. Verifying intron/exon borders using EST data. For the three family members that had EST data, the EST models were used in addition to plant intron/exon border consensus data to verify intron/exon borders in models.

For the two gene family members that had no known ESTs, the predicted models and the consensus data were the best available tools for resolving gene structure. The conceptually translated peptide sequences for the family members with ESTs were also uploaded as a comparison tool.

2.4.6.3. Identification of start codon through ORF (open reading frame) analysis. An ORF calculator was used to indicate the longest uninterrupted open reading frame that a model can produce. The ORF calculator was accessed through the APOLLO tool in DNA Subway.

2.4.7. Promoter Element Identification. Promoter elements, such as the TATA and CAAT boxes, were identified using PLACE (Plant Cis-Acting Regulatory DNA Elements). A sequence of ~1500 nucleic acids located upstream from each gene was used as an input. An overlap of at least 100 nucleotides at the 5' end of the gene was included as a means of measuring distance of elements from the start ATG. The "group by signal" option for results output was selected. The following information was collected and organized: TATA and CAAT box locations, distance of elements from each other in nucleotides, and distance of TATA elements from the start ATG.

The PLACE database was searched for elements associated with drought treatment and root hair in two separate searches. A record was made of the gene family members that contain an upstream element corresponding to the accession numbers that resulted from that search. In addition, all accession numbers were cross-referenced with gene family members according to the presence or absence of ESTs in an attempt to unveil expression patterns. Accession numbers corresponding to the following patterns were examined: those that belong to all genes with EST data, those

that belong to all genes without EST data, and those observed in all members of the family regardless of EST data. The information provided for the accession numbers meeting these criteria was analyzed for possible tissue-specific expression or environmental stress response expression patterns.

2.4.8. Evolutionary Analysis of Verified Gene Family Member Resolved Models. An evolutionary analysis of gene family members was conducted after model construction using a multiple alignment, codon alignment, phylogenetic tree construction, and were calculated using synonymous/non-synonymous substitution rates. In addition, pairwise alignments and dot plot comparisons were also utilized.

2.4.8.1. Multiple and pairwise alignments to analyze coding capacity and possible mutation sites. A multiple alignment of all gene family members was conducted using the conceptually translated peptide sequences of each in FASTA format as input. A second multiple alignment was carried out using the peptide sequence with the addition of conceptually translated amino acid residues on both the 5' and the 3' end of each sequence to make every sequence as long as the longest peptide sequence in the family.

A codon alignment was created using PAL2NAL. Protocol for the use of the PAL2NAL tool is outlined in Section 2.4.3.

Each gene family member was aligned pairwise against every other family member. Input data consisted of the coding nucleic acid sequence plus enough nucleotides extending from the 5' and 3' ends of the shorter sequence to make it as long as the longer sequence. The exonic regions of each gene in the output were delineated and the sequences examined for variants and/or mutations.

2.4.8.2. Generation of dot plots to assess similarity of sequence outside of the coding area. All gene family members were compared pairwise through a dot plot generator tool. The genomic nucleic acid sequence of each model in FASTA format was used as input. Parameters were set as follows: window size, 9; mismatch limit, 0. Additional pairwise comparisons were carried out between LJFgene3, LJFgene14, and LJFgnee1 to determine the level of sequence similarity of the regions of the chromosomes flanking these models. The sequences were extended up to 10kbp. The data generated from this tool was compared to the neighbor gene analysis.

2.5. NON-CODING SEQUENCE ANALYSIS

The conceptually translated peptide sequences of each non-coding sequence were placed into a web-based reverse translation tool (such as the Backtranseq tool supported by EMBL-EBI) to produce the corresponding nucleic acid sequences for use as a query for a BLASTn search against the *Glycine max* genome. A record was made of the details of the BLAST results including percent identity to gene family members, the range of nucleotides of the family members that the non-coding sequences corresponded to, and the composition of any other segments of DNA that the non-coding sequences correspond to outside of the gene family.

LJFnm19, LJFnm11, and LJFnm12 were compared to LJFgene3 (the basis for comparison for the family) using a dot plot matrix. A segment of nucleic acid sequence from chromosomes 19, 11, and 12 that included LJFnm19, LJfnm11, and LJfnm12, respectively, and a specified number of flanking nucleotides (enough to extend each side of the non-coding sequence to equal the length of LJFgene3 in

sequence) was dumped to FASTA format and used as input in a dot plot generator. Parameters were set as outlined in Section 2.4.7.3.

Non-coding nucleic acid sequences of LJFnm19, LJFnm11, and LJFnm12 were compared in a multiple alignment to LJFgene3, LJFgene14, and LJFgene1 genomic sequences. LJFnm19, LJFnm11, and LJFnm12 were also aligned against each other using conceptually translated peptide sequence, coding sequences, and genomic sequence as input for multiple alignments. All three non-coding nucleic acid sequences were individually plotted against their reverse compliment in a dot lot matrix to determine if dyad symmetry exists within the sequences.

The nucleic acid sequences of LJFnm19, LJFnm11, and LJFnm12 were submitted for analysis through FOMmiR, a web-based prediction tool that uses a fixed-order Markov model and is based on secondary structure.

A roughly 1kbp segment of DNA flanking the 5' side of LJFnm19, LJFnm11, and LJFnm12 was submitted for a signal scan through PLACE to determine the presence of any promoter elements. The query sequences were the nucleic acid sequence upstream of LJFnm19, LJFnm11, and LJFnm12 in FASTA format containing the LJFnm non-coding sequence as a location reference.

2.6. FUNCTIONAL ANALYSIS

SMART was used to search for elements that are already known to be in proteins such as domains of similar organization or composition, homologs of known structure, or signal sequences. The conceptually translated peptide sequences of all gene family members were input independently. The following search options were

selected: outlier homologs and homologs of known structures, PFAM domains, signal peptides, and internal repeats. The initial MEME results were obtained through search parameters that limited the width of a motif to 50 amino acids. A second MEME search was executed with the maximum width parameter increased to 100.

A subcellular localization prediction tool, CELLO v.2.5, was used to infer protein localization in the cell after synthesis through the analysis of the peptide sequence produced by a gene. A Prosite search was conducted to compare the sequence of each peptide against its collection of known sequence patterns with functional annotations. All peptide sequences were input as a single file and the option to exclude high occurrence motifs was unselected.

Secondary structure and fold prediction was performed using PSIPred, pGenTHREADER, and pDomTHREADER, protein structure prediction tools hosted by the bioinformatics resource ExPASy. All three tools can be accessed within the PSIPred protein sequence analysis workbench and ran simultaneously through a single input of gene family members as conceptually translated peptide sequences in FASTA format. Secondary structure and fold prediction was also performed using I-TASSER. Only the conceptually translated peptide sequence of the gene family member residing on chromosome 3 was used as input.

A hydropathy analysis was conducted to research the possibility of the gene family protein being a membrane transport protein or integral membrane protein. The hydropathy plotting system utilized for this research effort was Protscale, a tool within ExPASy. The peptide sequence of LJFgene3 (the family standard) was submitted twice for hydropathicity analysis using the Kyte-Doolittle scale, once with window

size 19 and once with window size 9. A window size of 19 provides a good means of determining whether the protein has transmembrane segments. A window size of 9 is used to determine hydrophobic versus hydrophilic regions of the protein as an indication of surface regions on a globular protein.

3. RESULTS

3.1. CHOICE OF GENE FAMILY AND IDENTIFICATION OF MEMBERS

3.1.1. Criteria Match. Table 3.1 provides a record of PFAM gene families meeting three of the criteria used to determine family of study: size of family <10, unknown function, and expansion in *Glycine max* relative to *Vitis vinifera* (Vvi. The common grape), *Populus trichocarpa* (Ptr, cottonwood or poplar tree), *Medicago truncatula* (Mtr, barrel medic), *Arabidopsis thaliana* (Ath, thale cress), *Arabidopsis lyrata* (Aly, rock cress), *Carica papaya* (Cpa, papaya or pawpaw), *Sorghum bicolor* (Sbi, sorghum), *Zea mays* (Zma, corn), *Brachipodium distachyon* (Bdi, purple false brome), and *Oryza sativa* (Osa, rice). The The LJFgene family data is identified by bold type.

The data from the PFAM database indicates that family PF07386 contains 4 putative genes. The source of this number is unpublished data released with an earlier version of the genome sequence for *Glycine max*. Generation of this figure took place through the use of predictive-algorithms that compared the sequence of the *Glycine max* genome to PFAM domains. The data from Supplimentary Table 5 of the Schmutz et al. publication [9] indicates that this family contains 3 putative genes. Data from this source was generated using a Phytozome clustering algorithm that analyzes syntenic regions between and within species for evidence of orthologs or paralogs that will indicate duplications.

Table 3.1. Record of PFAM families meeting criteria.

PFAM family	# genes via PFAM	PFAM functional annotation	*Glycine max* gene (SupTab 5)	Vvi:Ptr:Mtr:Gma:Ath:Aly:Cpa:Sbi:Zma:Bdi:Osa
PF06376	10	Protein of unknown function (DUF1070)	Glyma06g13060.1	0:4:1:8:2:2:2:0:0:0:0
PF06219	9	Protein of unknown function (DUF1005)	Glyma01g00910.1	1:0:1:3:1:1:1:1:1:1:1
PF07939	7	Protein of unknown function (DUF1685)	Glyma07g34280.1	0:0:1:2:0:0:0:0:0:0:0
PF06258	6	Protein of unknown function (DUF1022)	Glyma11g18750.1	2:3:0:5:2:2:2:2:2:2:2
PF06592	4	Protein of unknown function (DUF1138)	Glyma08g24930.1	0:1:0:4:1:1:0:0:0:0:0
PF07386	**4**	**Protein of unknown function (DUF1499)**	**Glyma08g39410.1**	**1:1:1:3:1:1:1:1:0:1:1**

3.1.2. Association Map Created Using BLAST within *Glycine max* Genome Browser.

An association map was created to determine whether all of the resultant putative genes are connected to each other based on similarity of sequence. Each gene model that results from a BLAST search that is a hit for that query is associated with that query, and this relationship can be represented by a connecting line. An association map of the gene family in this study can be seen in Figure 3.1. If every

gene model displays a connection to every other model, it provides strong support that those models belong to the same gene family.

LJFgene8	LJFgene1	LJFgene3	LJFgene14	LJFgene9
LJFgene1	LJFgene8	LJFgene8	LJFgene8	LJFgene8
LJFgene3	LJFgene3	LJFgene1	LJFgene1	LJFgene1
LJFgene14	LJFgene14	LJFgene14	LJFgene3	LJFgene3
LJFgene9	LJFgene9	LJFgene9	LJFgene9	LJFgene14
LJFnm19	LJFnm19	LJFnm19	LJFnm19	
LJFnm12	LJFnm12	LJFnm12	LJFnm12	
LJFnm11	LJFnm11	LJFnm11	LJFnm11	

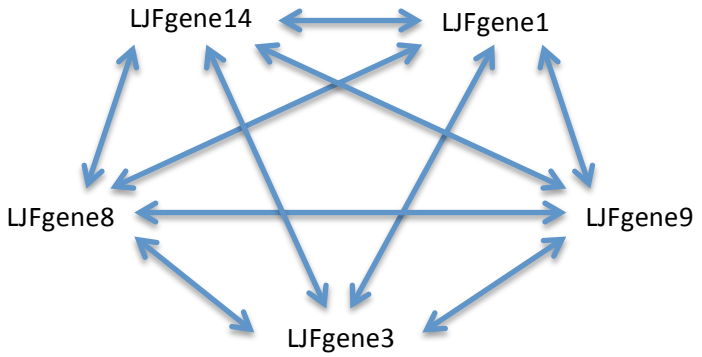


Figure 3.1. BLAST results and association map of the LJFgene family. Only predicted genes were included in the map. The non-coding sequences were omitted.

Based on sequence similarity indicated by BLAST searches, this family is comprised of five protein-coding genes (gene models predicted by algorithms) and potentially 3 non-coding sequences in *Glycine max*. The gene addresses and physical locations of putative gene family members are as follows:

- LJFgene3; sequence spanning from nucleotides 2524864 – 2528471 on chromosome 3.

- LJFgene14; sequence spanning from nucleotides 5788441 – 5792004 on chromosome 14.
- LJFgene1; sequence spanning from nucleotides 46754478 – 46758976 on chromosome 1.
- LJFgene8; sequence spanning from nucleotides 38952021 – 38955976 on chromosome 8.
- LJFgene9; sequence spanning from nucleotides 45453546 – 45455058 on chromosome 9.

Gene addresses are depicted graphically relative to position on the chromosome in Figure 3.2.

The number of genes in this *Glycine max* family is expanded 5:1 compared to the orthologous families in the nine other plant species (Vvi, Ptr, Mtr, Ath, Aly, Cpa, Sbi, Zma, Bdi, Osa) compared in Supplementary Table 5 [9].

3.1.3. Chromosome Maps. Figure 3.2 depicts the relative position of LJFgene family members on their respective chromosomes according to gene address.

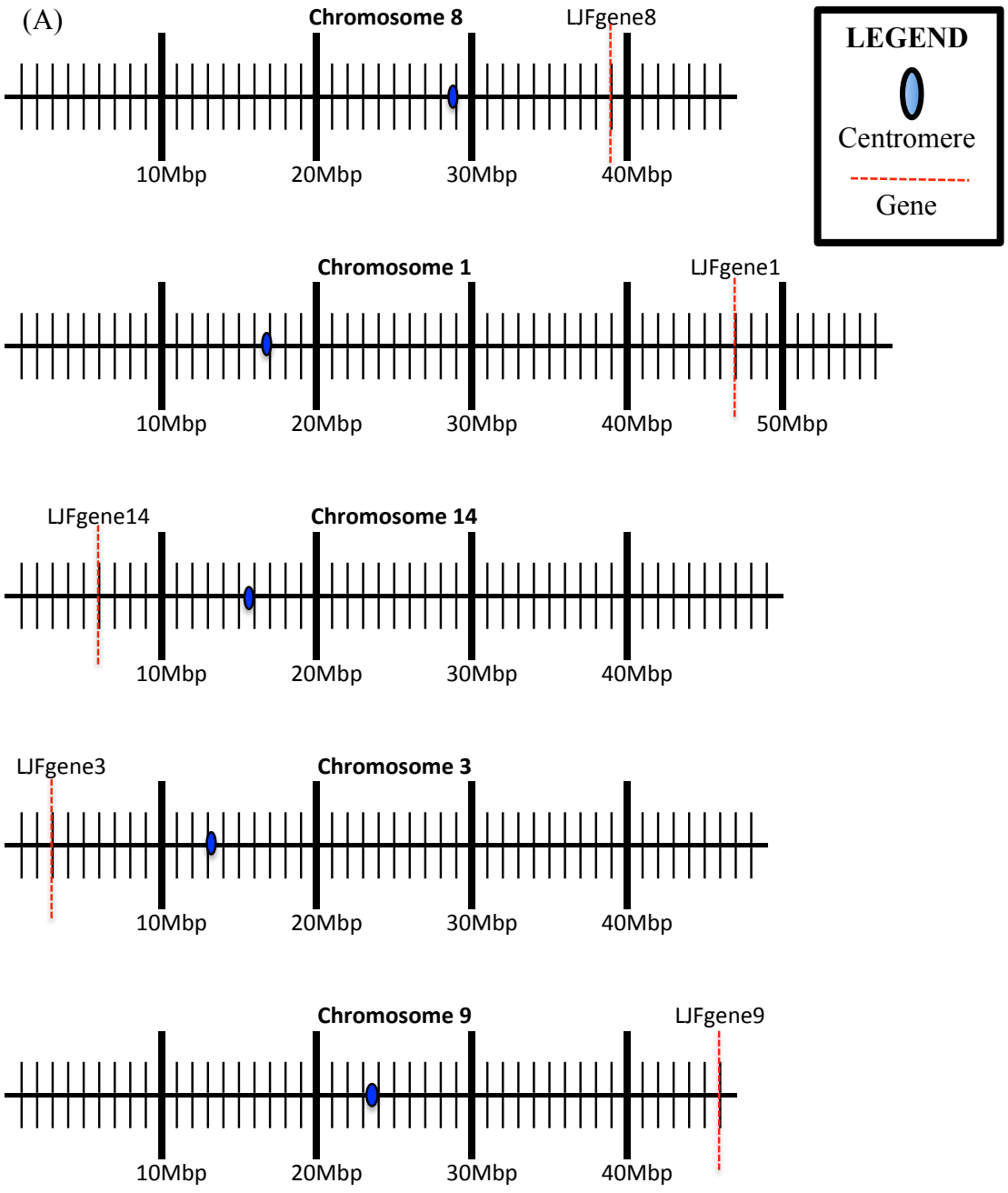


Figure 3.2. Chromosome maps. (A) Verified gene family members. (B) Non-coding sequences associated with the LJFgene family.

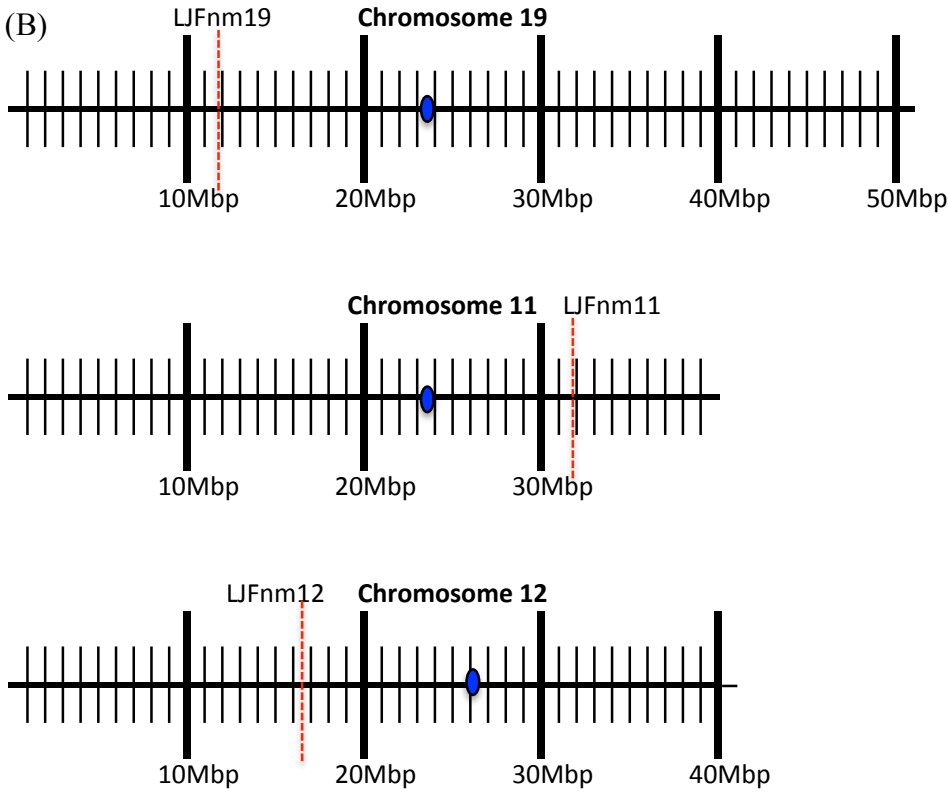


Figure 3.2. Chromosome maps. (CONT.)

3.1.4. General Summary of LJFgene Family Member Composition. Table 3.2 contains a record of the number of exons and introns in each LJFgene family member. It also contains a record of the number of residues in the genomic, coding, and conceptually translated peptide sequences, as well as the number of ESTs associated with each LJFgene family member.

Table 3.2. LJFgene family summary.

	Exons	Introns	Genomic sequence	Coding sequence	Amino acid sequence	ESTs
LJFgene3	7	6	3396	708	236	7
LJFgene14	6	5	3300	489	163	3
LJFgene1	7	6	4281	663	221	0
LJFgene8	6	5	3621	639	213	0
LJFgene9	4	3	1539	390	130	3

3.2. GENE STRUCTURE PREDICTION AND EST EXPRESSION ANALYSIS

3.2.1. Constructed Gene Models. Gene models were constructed using multiple data sources including algorithm-predicted models from FgenesH, Augustus, and GenomeScan, in addition to ESTs, and intron/exon border consensus data. Figure 3.3 illustrates exon placement and key intragenic coding signals along the genomic sequence corresponding to each LJFgene family member.

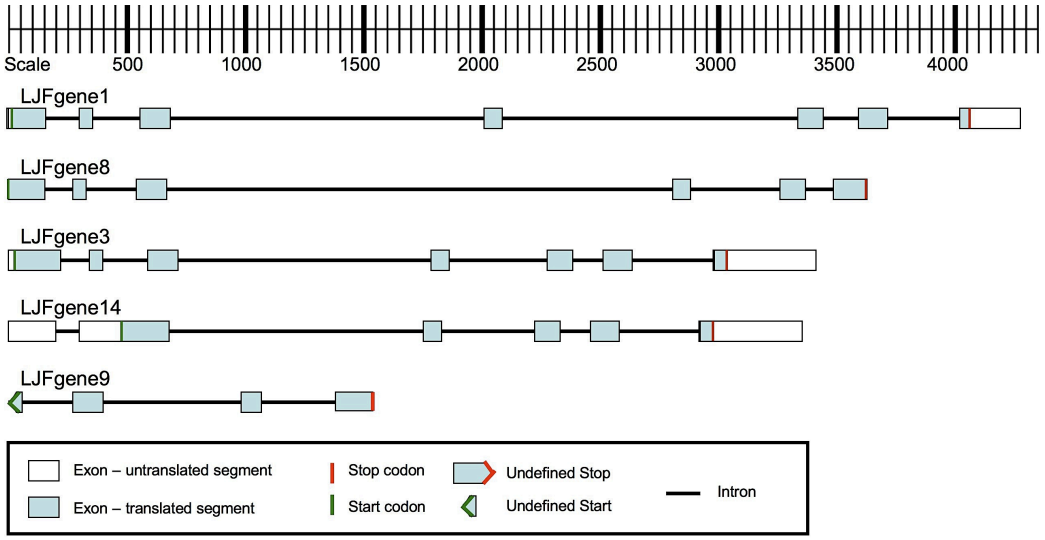


Figure 3.3. LJFgene models. (The scale is in base pair units.)

3.2.2. Decorated Sequences. Decorated sequences are viewable in Appendix E.

3.2.3. Promoter Element Locations. Well-known promoter elements, TATA and CAAT sequences, were predicted by PLACE. The program predicted the presence of these elements within a 1500 nucleic acid input sequence upstream of (and containing a partial exon 1 sequence) each LJFgene family member. The predicted sequences and their relative distances from one another, as well as from the start ATG, is recorded in Table 3.3.

Table 3.3. Promoter element locations and relative distances.

Promoter Elements					
TATA box	loc	CAAT box	loc	distance b/w elements	distance of TATA fr ATG
03g					
TB5-TTATTT	1042	CAAT	987	55	478
TB2-TATAAAT	1493	CCAAT	1163	330	27
TB5-TTATTT	743	CAAT	441	302	777
14g					
TB5-TTATTT	1413	CAAT	1279	134	9
TB2-TATAAAT	971	CAAT	815	156	451
TB5-TTATTT	946	CAAT	815	131	476
TB5-TTATTT	705	CAAT	618	87	717
01g					
TB5-TTATT	1491	CCAAT	1181	310	22
TB5-TTATT	1277	CCAAT	1181	96	236
TB5-TTATT	753	CAAT	504	249	760
TB5-TTATT	759	CAAT	504	255	754
08g					
TB4-TTTATATA	1439	CAAT	1424	15	62
TB5-TTATTT	1228	CAAT	1078	150	273
TB2-TATAAAT	1206	CAAT	1078	128	295
TB5-TTATTT	1179	CAAT	1078	101	322
TB5-TTATTT	1179	CAAT	1020	159	322
TB3-TATTAAT	1092	CAAT	1020	72	409

Table 3.3. Promoter element locations and relative distances. (CONT.)

09g					
TB2-TATAAAT	986	CAAT	945	41	515
TB4-TTTATATA	984	CAAT	945	39	517
TB5-TTATT	880	CAAT	569	311	621
TBSPAL	559	CAAT	355	204	942
TB5	552	CAAT	355	197	949
TB5	410	CAAT	300	110	1091
TB5	410	CAAT	294	116	1091
TB2-TATAAAT	343	CAAT	294	49	1158
TB2-TATAAAT	343	CAAT	196	147	1158
TB2-TATAAAT	343	CAAT	141	202	1158

3.2.4. EST Data for Intron/Exon Border Verification. The coding regions were determined using EST sequences where possible. Otherwise, models predicted by Augustus and FgenesH were used. The EST sequences for gene family members can be located in Appendix F. The NCBI accession numbers that correspond to the ESTs are listed below in Table 3.4. Accession numbers are linked to additional information about each EST located in the NCBI database. This information is also available in Appendix F.

Table 3.4. LJFgene family EST accession numbers and alignment scores.

LJFgene3		LJFgene14		LJFgene8	LJFgene9		LJFgene1
AI938349.1	201	EV264523.1	214		CD399608.1	72	
BE610616.1	409	FG993792.1	218		CF921901.1	247	
BM176973.1	398	HO044862.1	175		CF923165.1	133	
BM886799.1	288						
CA800126.1	457						
CO983876.1	387						
EV275072.1	456						

3.3. EVOLUTIONARY ANALYSIS

3.3.1. Neighbor Gene Analysis. Figure 3.4 is a spatial representation of both confirmed and predicted gene models flanking LJFgene family members (including both coding sequences and non-coding sequences) on the 5’ and 3’ sides. Each column represents a linear chromosome, each row represents a 5kbp segment of sequence, and colored blocks represent genes. The genes are color coded according to their function. The full function can be found in the functional annotation color key. A condensation of Figure 3.4A was created by removal of extragenic spaces for better pattern recognition in consideration of the evolution of the family and is illustrated in Figure 3.4B.

(A)

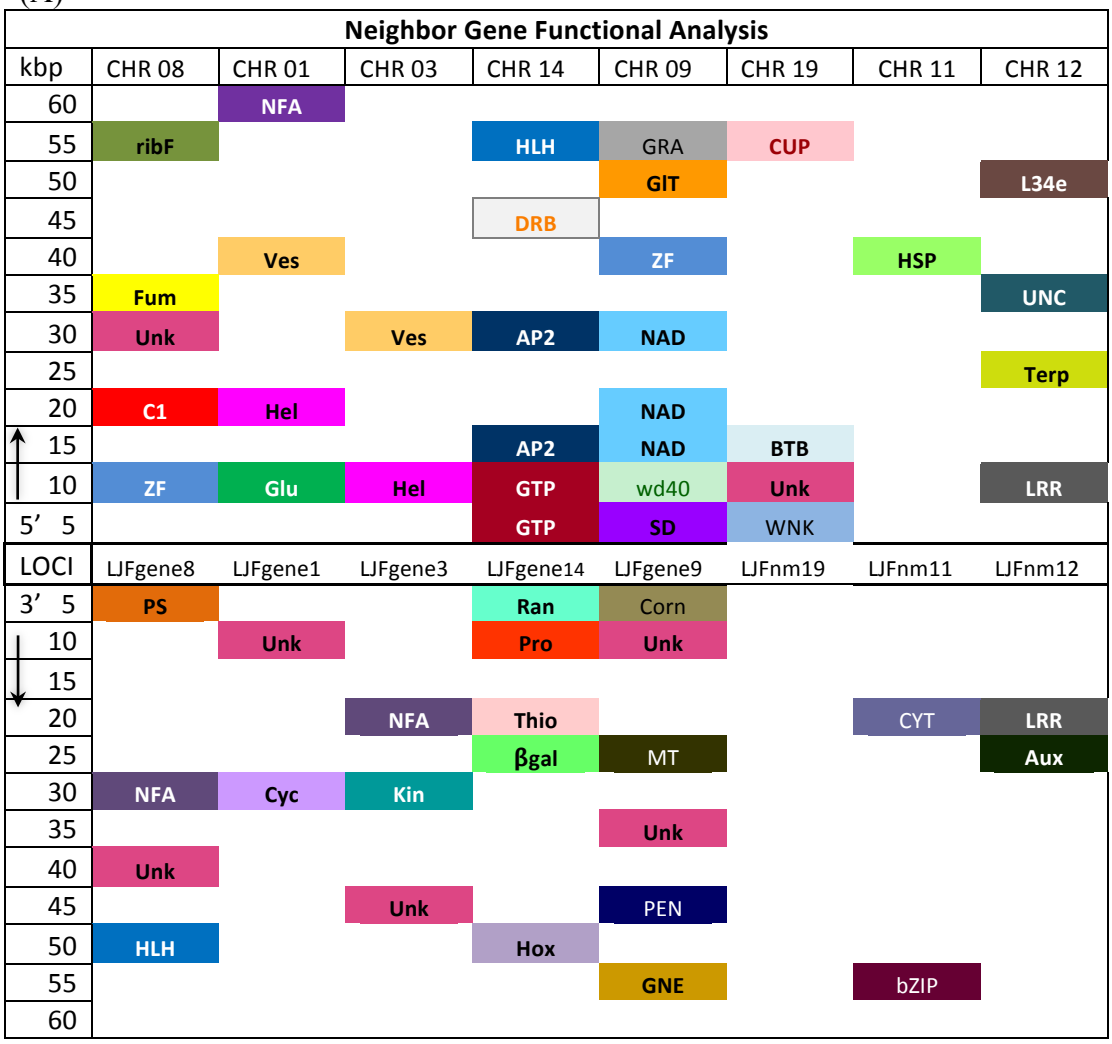


Figure 3.4. (A) Neighbor gene functional analysis. (B) Condensation of Neighbor gene functional analysis.

Functional Annotation Color Key			
ZF	Zinc finger	GRA	GRAS family transcription factor
C1	C1domain/Thioredoxin/nucleoredoxin	Corn	Cornichon protein
Unk	unknown	MT	Microtubule-assoc protein-anaphase
Fum	Fumble, pantothenate kinase,	PEN	PENTATRICOPEPTIDE REPEAT
ribF	GTP-binding ADP-ribosylation factor	GNE	Guanine nucleotide exchange factor
PS	Proteasome subunit	BTB	BTB/POZ domain
NFA	no functional annotations	CUP	Cupin domain
HLH	Helix-loop-helix DNA-binding	CYT	CYTOCHROME
Glu	Glutathione S-transferase	bZIP	bZIP transcription factor
Hel	Helicase	HSP	SMALL HEAT-SHOCK PROTEIN
Ves	Vesicle trafficking	LRR	Leucine Rich Repeat
Cyc	RNA 3'-terminal phosphate cyclase	Terp	Terpene synthase
Kin	kinase (leucine rich repeat)	UNC	uncharacterized
GTP	GTP binding	WNK	SERINE/THREONINE-PROTEIN KINASE
AP2	AP2; transcription factor		WNK (WITH NO LYSINE)-RELATED
DRB	double-stranded RNA binding	Aux	Auxin responsive protein
Ran	Ran binding protein	L34e	Ribosomal protein L34e
Pro	ATP-dependent protease/peptidase		
Thio	Thioredoxin		
βgal	Beta-galactosidase		
Hox	Homeobox domain assoc w/ HOX		
SD	STEROL DESATURASE		
wd40	WD40 repeat protein		
NAD	NAD dependent epimerase		
GlT	Glycosyl transferase		

Figure 3.4. (A) Neighbor gene functional analysis. (B) Condensation of Neighbor gene functional analysis. (CONT.)

(B)

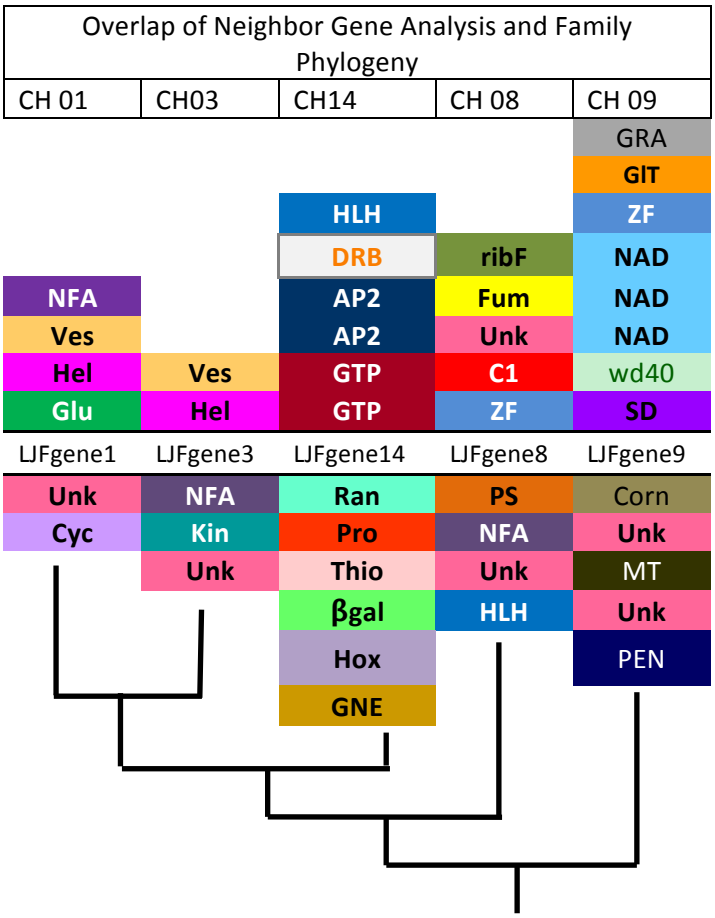


Figure 3.4. (A) Neighbor gene functional analysis. (B) Condensation of Neighbor gene functional analysis. (CONT.)

3.3.2. Synonymous and Nonsynonymous Substitution Rates. Table 3.5 contains the calculations for synonymous and nonsynonymous substitutions for each possible pairwise comparison between LJFgene family members. Table 3.6 contains the calculations for synonymous and nonsynonymous substitutions between LJFgene family members and orthologous genes found in the species *P. vulgaris, M. truncatula, A. thaliana, O. sativa, S. moellendorffii, P. patens, V. carteri,* and *C. reinhardtii*, as well as between the orthologous species' genes.

Table 3.5. Synonymous and non-synonymous calculations for LJFgene family.

Seq's compared		Sd	Sn	S	N	ps	pn	ds	dn	ds/dn	ps/pn
LJFgene3	LJFgene14	19.0000	23.0000	106.6667	382.3333	0.1781	0.0602	0.2034	0.0627	3.2431	2.9610
LJFgene3	LJFgene1	18.5000	9.5000	142.0000	512.0000	0.1303	0.0186	0.1431	0.0188	7.6169	7.0215
LJFgene3	LJFgene8	40.8333	64.1667	136.5000	490.5000	0.2991	0.1308	0.3817	0.1438	2.6552	2.2867
LJFgene3	LJFgene9	46.6667	70.3333	84.6667	302.3333	0.5512	0.2326	0.9958	0.2785	3.5755	2.3693
LJFgene14	LJFgene1	15.0000	23.0000	105.8333	383.1667	0.1417	0.0600	0.1571	0.0626	2.5109	2.3612
LJFgene14	LJFgene8	26.0000	51.0000	102.5000	368.5000	0.2537	0.1384	0.3096	0.1530	2.0236	1.8328
LJFgene14	LJFgene9	49.3333	84.6667	83.3333	303.6667	0.5920	0.2788	1.1681	0.3486	3.3507	2.1233
LJFgene1	LJFgene8	37.1667	68.8333	138.5000	497.5000	0.2684	0.1384	0.3321	0.1529	2.1717	1.9395
LJFgene1	LJFgene9	44.6667	69.3333	84.1667	302.8333	0.5307	0.2289	0.9222	0.2732	3.3759	2.3180
LJFgene8	LJFgene9	45.0000	70.0000	84.1667	302.8333	0.5347	0.2312	0.9359	0.2763	3.3866	2.3130

*The column highlighted grey is data of interest for generating phylogenetic models.

Table 3.6. Ortholog synonymous substitutions.

Compare	vs	ps	ps/pn	Statistics	
				Mean ps	Mean ps/pn
LJFgene3	LJFgene1	0.13	7.02	0.27	N/A
LJFgene3	LJFgene14	0.18	2.96		
LJFgene3	LJFgene8	0.23	2.28		
LJFgene3	LJFgene9	0.55	1.82		
LJFgene1	LJFgene14	0.14	2.36	0.3	N/A
LJFgene1	LJFgene8	0.19	1.89		
LJFgene1	LJFgene9	0.58	1.97		
LJFgene14	LJFgene8	0.25	1.83	0.42	N/A
LJFgene14	LJFgene9	0.58	1.66		
LJFgene8	LJFgene9	0.57	1.91	N/A	N/A
LJFgene3	Pvu	0.29	3.17	0.33	2.83
LJFgene1	Pvu	0.27	3.38		
LJFgene14	Pvu	0.27	3.47		
LJFgene8	Pvu	0.30	2.14		
LJFgene9	Pvu	0.55	1.97		
LJFgene3	Mtr	0.40	3.30	0.44	3.04
LJFgene1	Mtr	0.39	3.55		
LJFgene14	Mtr	0.41	3.88		
LJFgene8	Mtr	0.41	2.44		
LJFgene9	Mtr	0.62	2.13		
Pvu	Mtr	0.40	2.94		
LJFgene3	Ath	0.70	2.91	0.7	2.85
LJFgene1	Ath	0.66	2.75		
LJFgene14	Ath	0.72	4.18		
LJFgene8	Ath	0.70	2.45		

Table 3.6. Ortholog synonymous substitutions. (CONT.)

LJFgene9	Ath	0.75	2.12		
Mtr	Ath	0.71	2.77		
Pvu	Ath	0.66	2.78		
LJFgene3	Osa	0.79	3.00	0.82	2.96
LJFgene1	Osa	0.81	3.09		
LJFgene14	Osa	0.84	3.68		
LJFgene8	Osa	0.81	2.69		
LJFgene9	Osa	0.76	2.03		
Pvu	Osa	0.84	3.05		
Mtr	Osa	0.88	3.16		
Ath	Osa	0.86	2.97		
Smo	LJFgene3	0.83	3.96	0.86	3.69
Smo	LJFgene1	0.88	4.17		
Smo	LJFgene14	0.88	4.34		
Smo	LJFgene8	0.89	3.55		
Smo	LJFgene9	0.85	2.11		
Smo	Pvu	0.87	4.09		
Smo	Mtr	0.82	3.93		
Smo	Ath	0.86	3.64		
Smo	Osa	0.84	3.38		
LJFgene3	Ppa13v6	0.83	1.33	0.82	1.30
LJFgene1	Ppa13v6	0.85	1.34		
LJFgene14	Ppa13v6	0.81	1.33		
LJFgene8	Ppa13v6	0.82	1.28		
LJFgene9	Ppa13v6	0.77	1.06		
Pvu	Ppa13v6	0.82	1.30		
Mtr	Ppa13v6	0.84	1.35		
Ath	Ppa13v6	0.83	1.30		
Osa	Ppa13v6	0.83	1.35		
Smo	Ppa13v6	0.81	1.36		
Ppa47v6	LJFgene3	0.74	2.12	0.77	2.41
Ppa47v6	LJFgene1	0.73	2.09		
Ppa47v6	LJFgene14	0.74	2.91		
Ppa47v6	LJFgene8	0.76	2.02		
Ppa47v6	LJFgene9	0.80	1.98		
Ppa47v6	Pvu	0.76	2.15		
Ppa47v6	Mtr	0.79	2.26		
Ppa47v6	Ath	0.79	2.19		
Ppa47v6	Osa	0.80	2.19		
Ppa47v6	Smo	0.86	5.47		
Ppa47v6	Ppa13v6	0.72	1.13		
LJFgene3	Vca	0.90	2.17	0.88	2.19
LJFgene1	Vca	0.90	2.13		
LJFgene14	Vca	0.98	2.83		
LJFgene9	Vca	0.90	1.82		
LJFgene8	Vca	0.87	1.93		
Pvu	Vca	0.94	2.32		
Mtr	Vca	0.90	2.10		
Ath	Vca	0.87	2.08		
Osa	Vca	0.73	1.73		
Smo	Vca	0.86	3.40		
Ppa13v6	Vca	0.78	1.30		
Ppa47v6	Vca	0.91	2.42		
LJFgene3	Cre	0.88	2.11	0.83	2.11
LJFgene1	Cre	0.89	2.12		
LJFgene14	Cre	0.91	2.62		
LJFgene8	Cre	0.89	1.93		

Table 3.6. Ortholog synonymous substitutions. (CONT.)

LJFgene9	Cre	0.97	1.90		
Pvu	Cre	0.86	2.07		
Mtr	Cre	0.91	2.14		
Ath	Cre	0.87	1.96		
Osa	Cre	0.69	1.61		
Smo	Cre	0.80	2.98		
Ppa13v6	Cre	0.73	1.21		
Ppa47v6	Cre	0.82	2.01		
Vca	Cre	0.55	2.72		

Code Key for Table 3.5 and Table 3.6:
Sd: number of observed synonymous substitutions
Sn: number of observed non-synonymous substitutions
S: number of potential synonymous substitutions
N: number of potential non-synonymous substitutions
ps: proportion of observed synonymous substitutions (Sd/S)
pn: proportion of observed non-synonymous substitutions (Sn/N)
ds: Jukes-Cantor correction for multiple ps
dn: Jukes-Cantor correction for multiple pn
ds/dn: ratio of synonymous to non-synonymous substitutions

3.3.3. Phylogenetic Trees. Figure 3.5 illustrates the phylogenetic trees generated for the LJFgene family members as well as a tree containing a broader diversity of plant species and the LJFgene family.

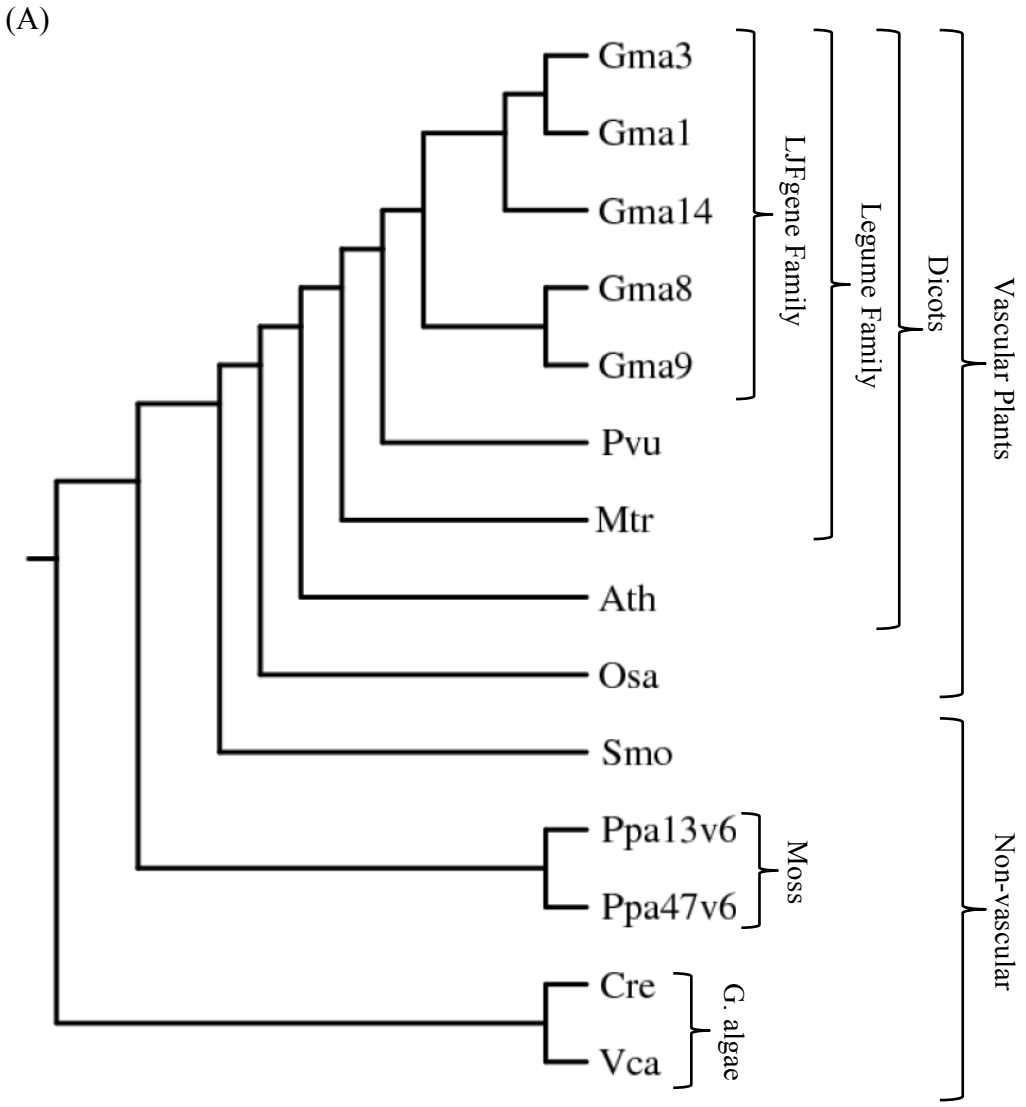


Figure 3.5. Phylogentic results. (A) Phylogenetic tree of diverse plant evolution (phenogram). (B) LJFgene family phylogenetic tree (unrooted radial display). (C) LJFgene family phenogram with corresponding gene model.

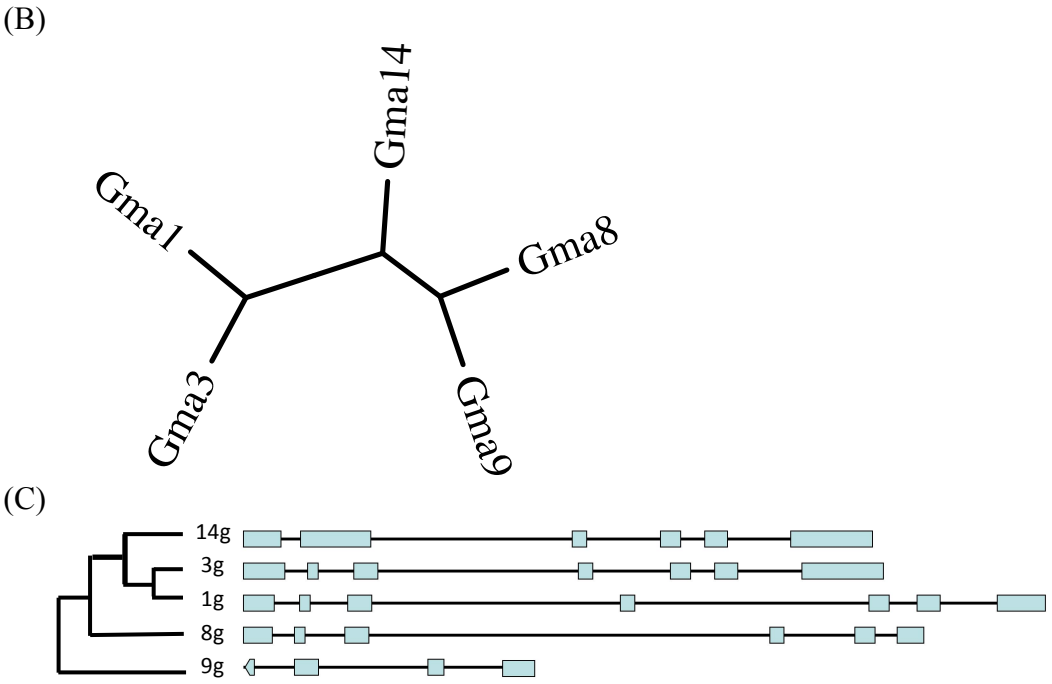


Figure 3.5. Phylogentic results. (CONT.)

Table 3.7 contains the gene addresses of predicted genes in plant species other than *Glycine max* that resulted as BLAST hits when the conceptually translated peptide sequence of each LJFgene family member was used as input for searches against each species' genome. The resulting genes are putative orthologs to the genes in the LJFgene family.

Table 3.7. BLAST hits in orthologous species.

Species	LJFgene3	LJFgene8	LJFgene1	LJFgene14	LJFgene9
M. truncatula	Medtr4g024760	Medtr4g024760	Medtr4g024760	Medtr4g024760	Medtr4g024760
P. vulgaris	phvulv091010655m	phvulv091010655m	phvulv091010655m	phvulv091010655m	phvulv091010655m
A. thaliana	AT3G60810	AT3G60810	AT3G60810	AT3G60810	AT3G60810
Z. mays	no transcript	no transcript	no transcript	no transcript	no transcript
O. sativa	LOC_Os03g64140	LOC_Os03g64140	LOC_Os03g64140	LOC_Os03g64140	LOC_Os03g64140
S. Moellendorffii	101401	101401	101401	101401	101401
P. patens	Pp1s10_47V6 (72)	Pp1s10_47V6 (70.1)	Pp1s10_47V6	Pp1s10_47V6	Pp1s10_47V6
	Pp1s223_13V6 (40.8)		Pp1s223_13V6	Pp1s223_13V6	
C. reinhardtii	Cre11.g468750	Cre11.g468750	Cre11.g468750	Cre11.g468750	Cre11.g468750
V. carterii	Vocar20006092m.g	Vocar20006092m	Vocar20006092m	Vocar20006092m	Vocar20006092m

3.3.4. Potential Coding Capacity. In order to determine whether the sequences beyond the coding regions of the genes that appear to produce a truncated product (LJFgene14, LJFgene1, LJFgene8, and LJFgene9) once contained coding capacity, multiple alignments were conducted using coding sequence plus 5' and 3' extended flanking sequences.

3.3.4.1. Multiple alignment of nucleic acid sequences. Figure 3.6 illustrates the multiple alignment of the nucleic acid sequences for LJFgene family members.

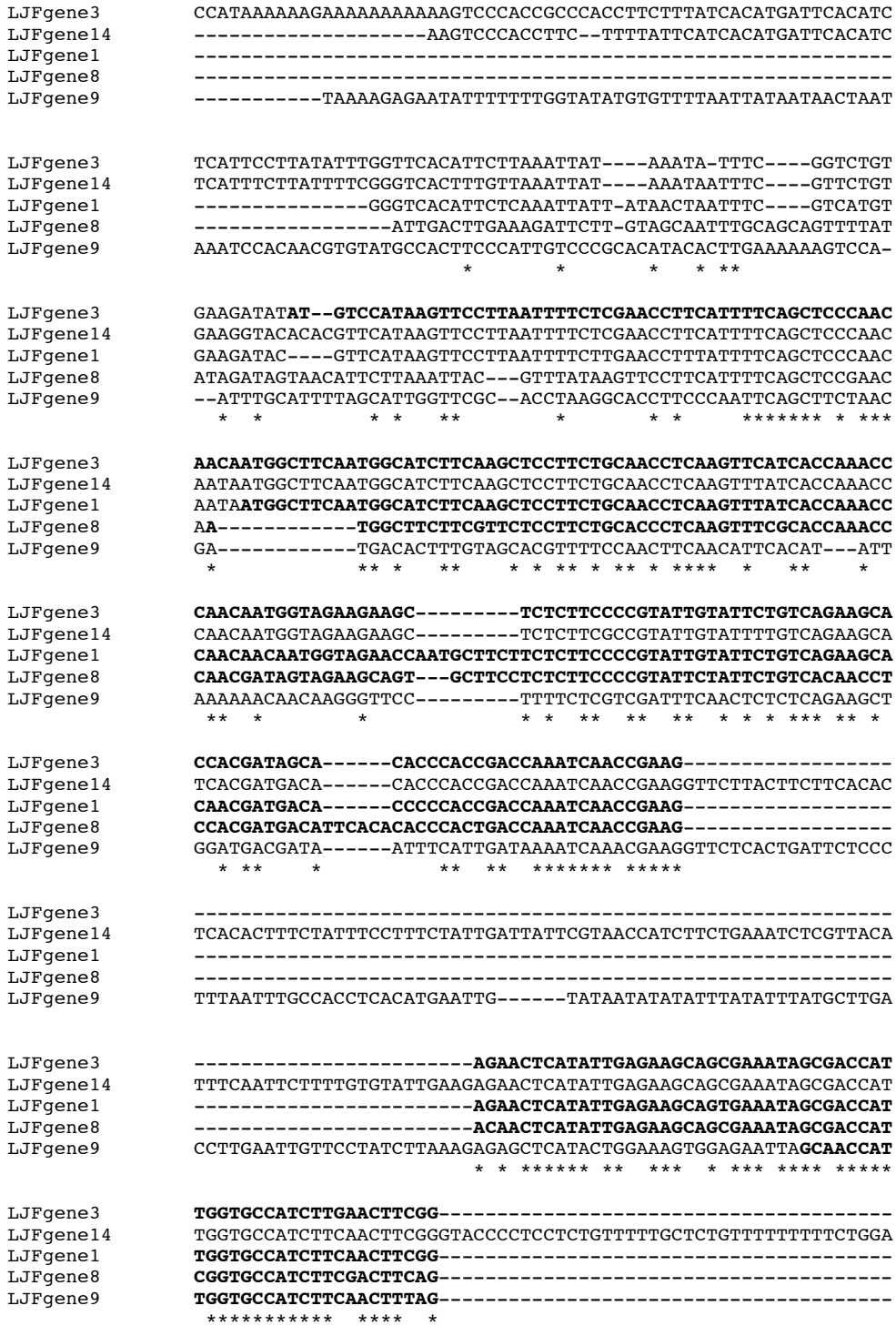


Figure 3.6. Multiple alignment of coding sequences (bold face type) of gene family members extended on both the 3' and 5' ends. Dashes (-) represent gaps in sequence and a star (*) below a column of nucleotides represents an identity match is present in all aligned sequences at that position, i.e. 100% conservation.

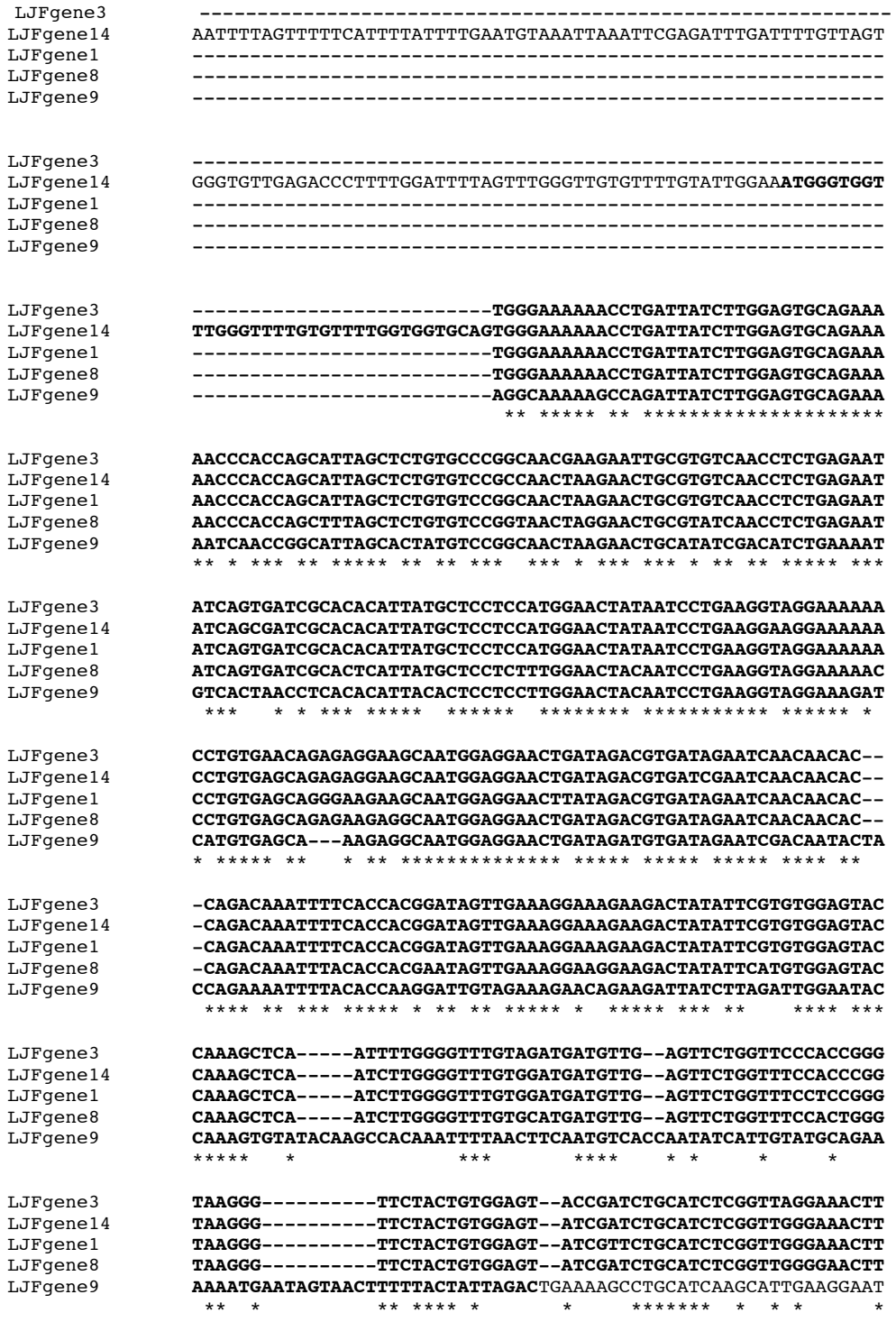


Figure 3.6. Multiple alignment of coding sequences (bold face type) of gene family members extended on both the 5' and 3' ends. (CONT.)

```
LJFgene3        TGATTTTGATGTGAACAGAAAAAGAATAAAGGCACTGCGACAAGAGTTGGAG-AAGAAAG
LJFgene14       TGATTTTGATGTGAACAGAAAAAGAATAAAGGCACTGAGACAAGAGTTGGAG-AAGAAAG
LJFgene1        TGATTTTGATGTGAACAGAAAAAGAATAAAGGCACTGAGACAAGAGTTGGAG-AAGAAAG
LJFgene8        TGATTTTGATGTGAATAAGAAAAGAATAAAGGTAT---------GTTTGTAT-CATTCCT
LJFgene9        GGAGTTTCTT-TGATCATTGGACTGCCCATCTCATAGTAACTCATCTTGAAGCAATTAAT
                 ** ***  * ***  *    *      *    *           *** *   *

LJFgene3        GATGGGCATCTCAAGACACCATATGATGAATAAACTCAGGCAGAATTAACATCAGCATCT
LJFgene14       GATGGACATCTCAAGATACCATATGATTAATAAACTCAGGCTGAATTAGCATCAGCATCT
LJFgene1        GATGGGCATCTCAAGACACCATATGATGAAAAAACTTAGGCAGAATTCACATCAGCATCT
LJFgene8        TTGTGCTGTCTCGGTAGTTAACATGAAGAAATGATTAAAAGATATTTTGT-----CCTTT
LJFgene9        GCAGTAAAAACTAACTCATCGTGAAAAGTTCATTCTCTGCTTTATTTAAATTTTTACAGC
                                       *         *       * **

LJFgene3        AAGCAAATATTATTTCATATACTTTGTGACCTTGTATACATTTGTA-TTAGATACAAA-T
LJFgene14       AAGCAAATATTATTTCATATACTTTG-GACCTTGTATACTTTTGTA-TTAGATACAAA-T
LJFgene1        AAGAAAATATTGTTTCATATACATTGTAACCTTGTATACTTTTGTA-TTAGATACAAAAT
LJFgene8        AGGTTT-TTTGGTTATATTTAGTTTG--ATTTTTTATTTTTTAAAAGTTAAATTTAGTCC
LJFgene9        AAGTAGATTGGAATGCATGGTTTTTGCCATGTTTTATACTTGACAAAGATAATGCAAACT
                * *    *     *  **     ***  *  ** ***   *    *     **  *

LJFgene3        CTCAC---
LJFgene14       CGCACA--
LJFgene1        CTCA----
LJFgene8        TT------
LJFgene9        ATAAACAC
```

Figure 3.6. Multiple alignment of coding sequences (bold face type) of gene family members extended on both the 5' and 3' ends. (CONT.)

3.3.4.2. Multiple alignment of conceptually translated peptide sequences.

Figure 3.7 illustrates the multiple alignment of the conceptually translated amino acid sequences for LJFgene family members.

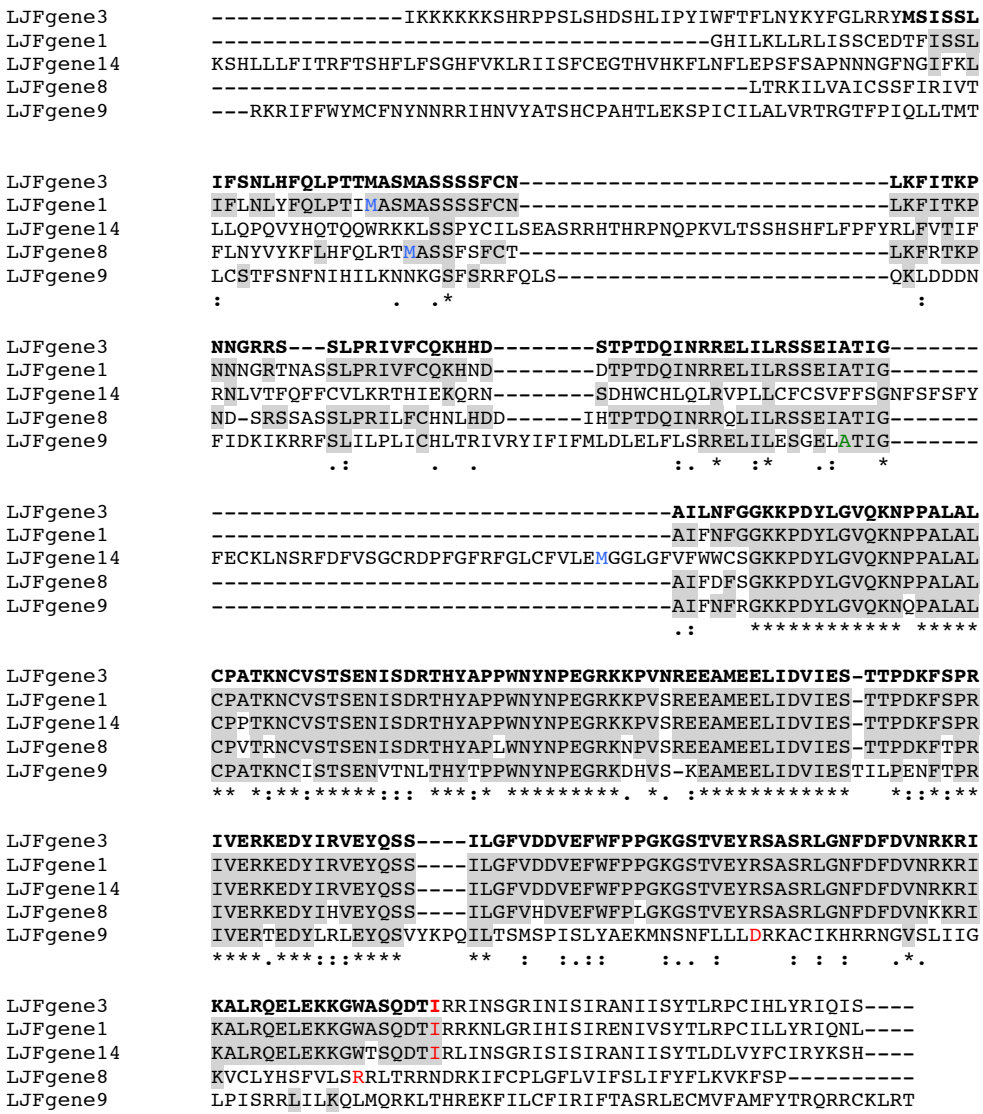


Figure 3.7. Multiple alignment of conceptually translated peptide sequences of gene family members extended on both the 3' and 5' ends. The sequence of LJFgene3 is indicated by bold face type. Highlighted residues in other LJFgene sequences indicate identity shared with LJFgene3. Residues colored blue (X) indicate the first predicted residue of a gene (with the exception of LJFgene3) and residues colored red (X) indicate the last predicted residue of a gene (with exception of LJFgene3). Dashes (-) between residues represent gaps in sequence; a star (*) below a column of residues represents an identity match is present in all aligned sequences at that position, i.e. 100% conservation; a colon (:) represents strong chemical property conservation between residues at a position (based on a scoring matrix threshold); a period (.) represents weak chemical property conservation between residues at a position (based on a scoring matrix threshold).

3.3.4.3. Codon alignment of gene family members extended on both the 3' and 5' ends.

Figure 3.8 illustrates the codon alignment for LJFgene family members.

Figure 3.8. Codon alignment with extended sequence. All intermittent stops in extended sequences were arbitrarily replaced with R (arginine residues) to extend the reading frame for acceptance by this program. Red residues represent replaced codons. Residues in bold type are representative of the coding sequence. Dashes (-) indicate gaps.

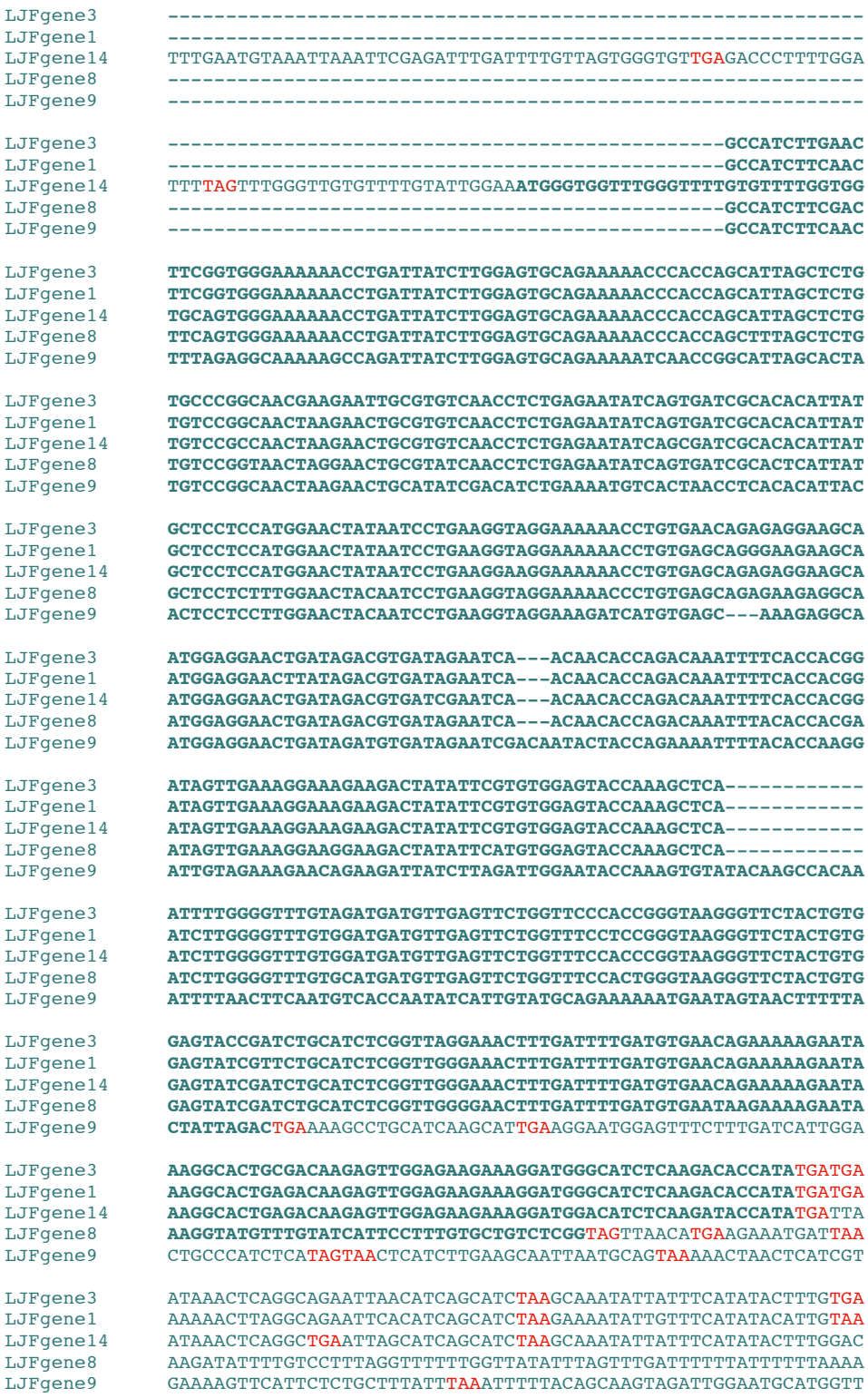


Figure 3.8. Codon alignment with extended sequence. (CONT.)

```
LJFgene3      CCTTGTATACATTTGTATTAGATACAAATCTCA------------
LJFgene1      CCTTGTATACTTTTGTATTAGATACAAAATCTC------------
LJFgene14     CTTGTATACTTTTGTATTAGATACAAATCGCAC------------
LJFgene8      GTTAAATTTAGTCCT------------------------------
LJFgene9      TTTGCCATGTTTTATACTTGACAAAGATAATGCAAACTATAAACA
```

Figure 3.8. Codon alignment with extended sequence. (CONT.)

3.3.4.4. Pairwise dot plot matrices. Pairwise dot plot matrices of gene family members provide an alternative method of determining sequence similarity both within and surrounding the predicted coding sequence. A dot plot matrix of LJFgene3 plotted against LJFgene14 is illustrated in Figure 3.9. A dot plot matrix of LJFgene3 plus extended sequence plotted against LJFgene14 plus extended sequence is illustrated in Figure 3.10. A dot plot matrix of LJFgene3 plus extended sequence plotted against LJFgene14 plus extended sequence and shifted for 3' boundary analysis is illustrated in Figure 3.11. Two dot plot matrices of LJFgene3 plus extended sequence plotted against LJFgene1 plus extended sequence are illustrated in Figures 3.12 and 3.13. An analysis of a 10 kbp sequence from chromosomes 3 and chromosome 1 beginning at the second to last exon of each gene model and extending past the 3' end and up to the most proximal 3' neighbor gene is illustrated in Figures 3.14, 3.15, and 3.16. The similarity of the sequences extending beyond the 3' ends of the gene family members does extend into the neighbor gene on chromosome 3, but not into the neighbor gene on chromosome 1.

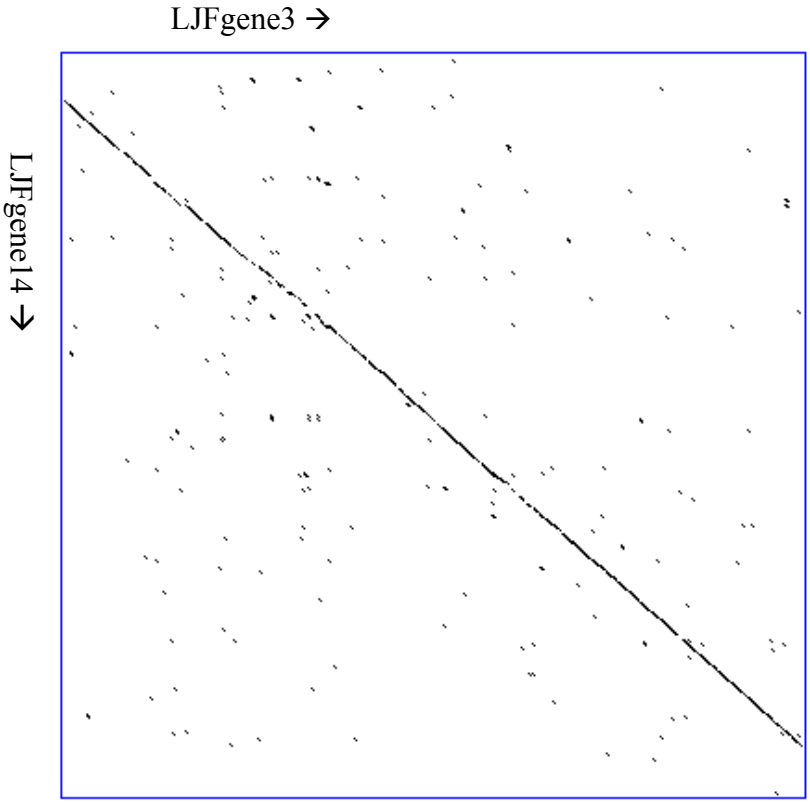


Figure 3.9. Dot plot: LJFgene3 (genomic sequence) vs LJFgene14 (genomic sequence). LJFgene3 on x-axis and LJFgene14 on y-axis.

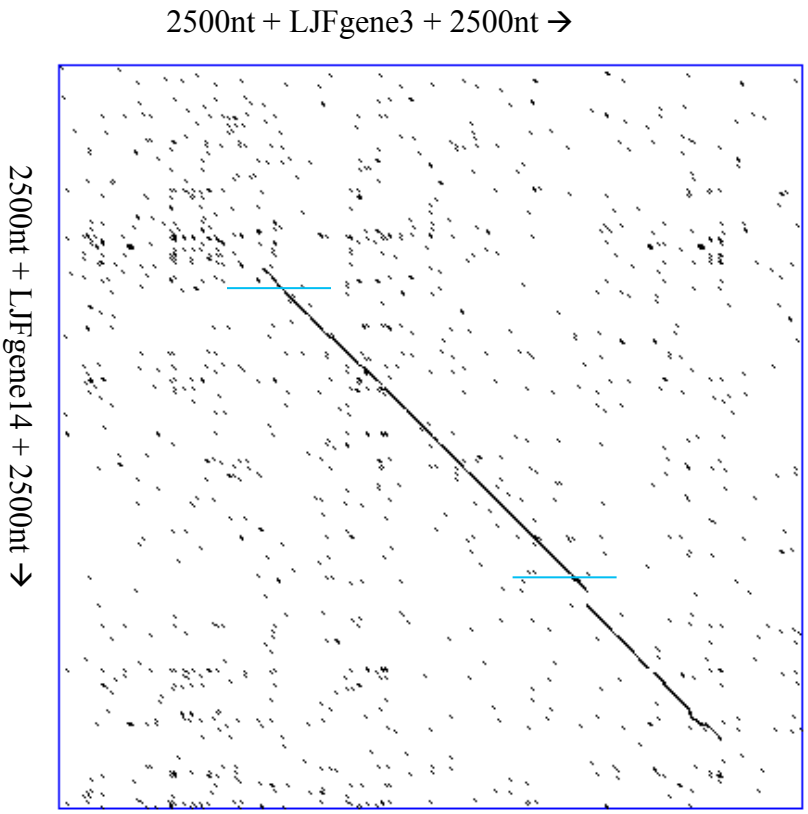


Figure 3.10. Dot plot: LJFgene3 genomic sequence plus approximately 2500nt extension from both 5' and 3' gene model boundaries (x-axis) vs. LJFgene14 genomic sequence plus approximately 2500nt extension from both 5' and 3' gene model boundaries (y-axis). Blue lines indicate boundary between genomic sequence identity and flanking sequences.

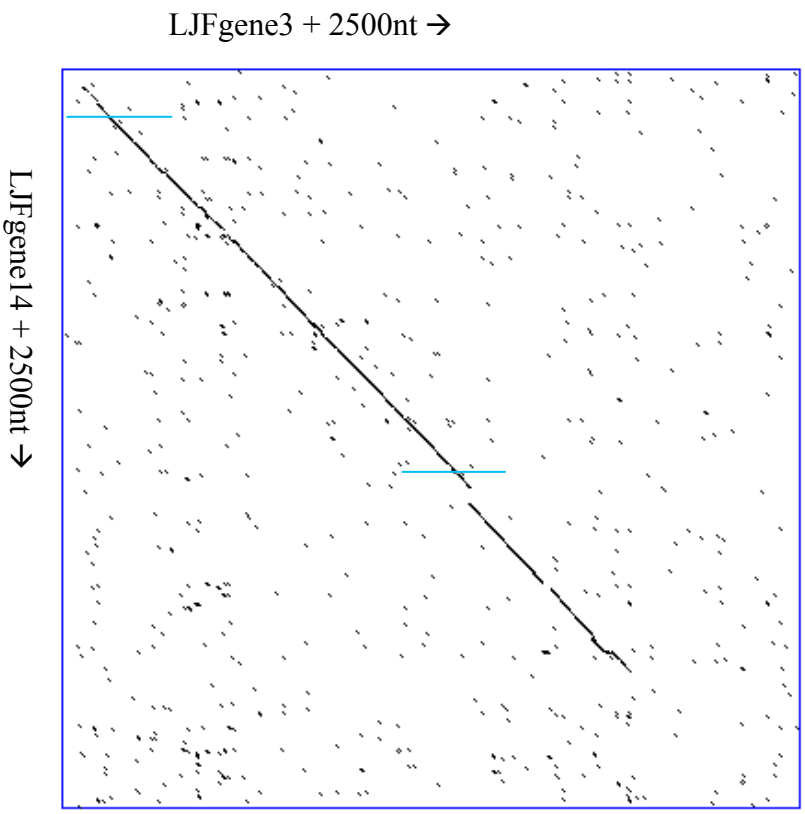


Figure 3.11. Dot plot: LJFgene3 genomic sequence plus approximately 2500nt extension (x-axis) vs. LJFgene14 genomic sequence plus approximately 2500nt extension (y-axis). Adjustment: 5'extention removed to shift plot for similarity analysis of sequence beyond 3' end of genes. Blue lines indicate boundary between genomic sequence identity and flanking sequences.

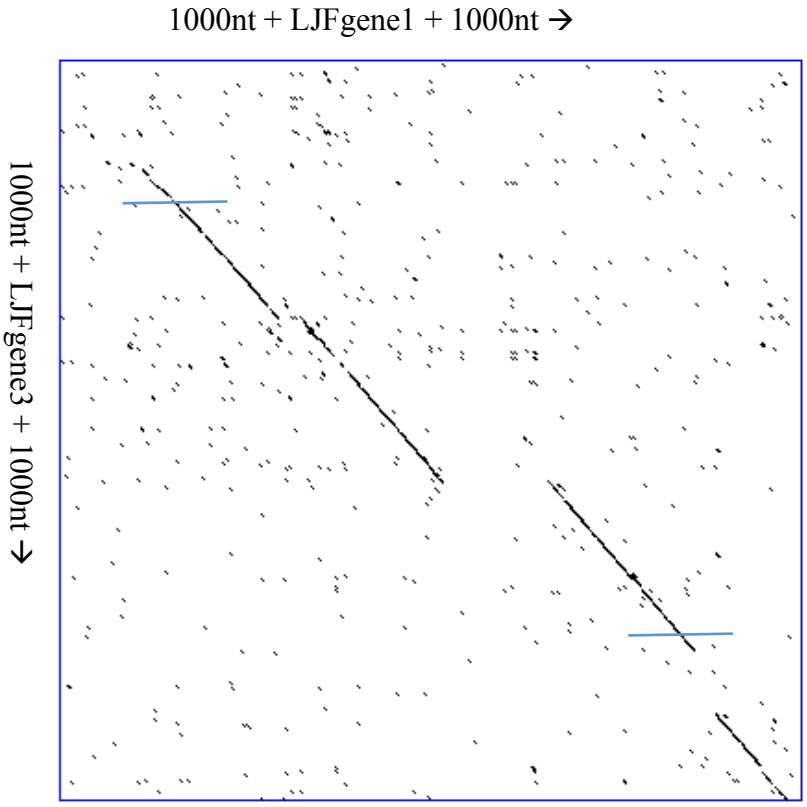


Figure 3.12. Dot plot: LJFgene1 genomic sequence plus 1000nt extension from both 5' and 3' gene model boundaries (x-axis) vs. LJFgene3 genomic sequence plus 1000nt extension from both 5' and 3' gene model boundaries (y-axis). Blue lines indicate boundary between genomic sequence identity and flanking sequences.

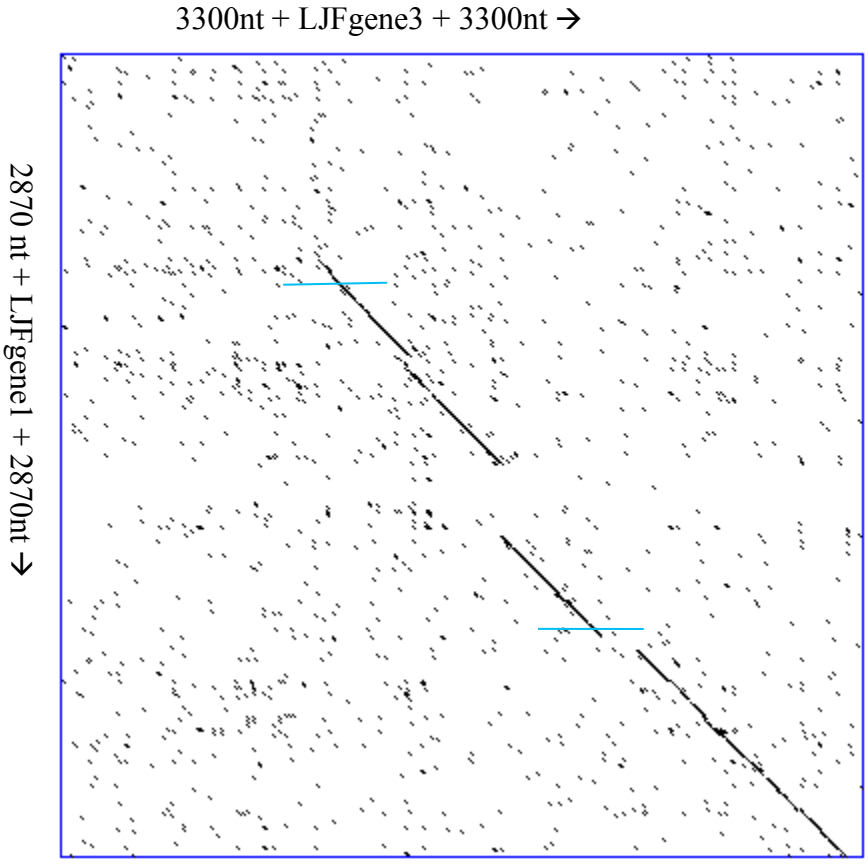


Figure 3.13. Dot plot: LJFgene3 genomic sequence plus approximately 3300 nucleotide extension from both 5' and 3' gene model boundaries (x-axis) vs. LJFgene1 genomic sequence plus approximately 2870 nucleotide extension from both 5' and 3' gene model boundaries (y-axis). Blue lines indicate boundary between genomic sequence identity and flanking sequences.

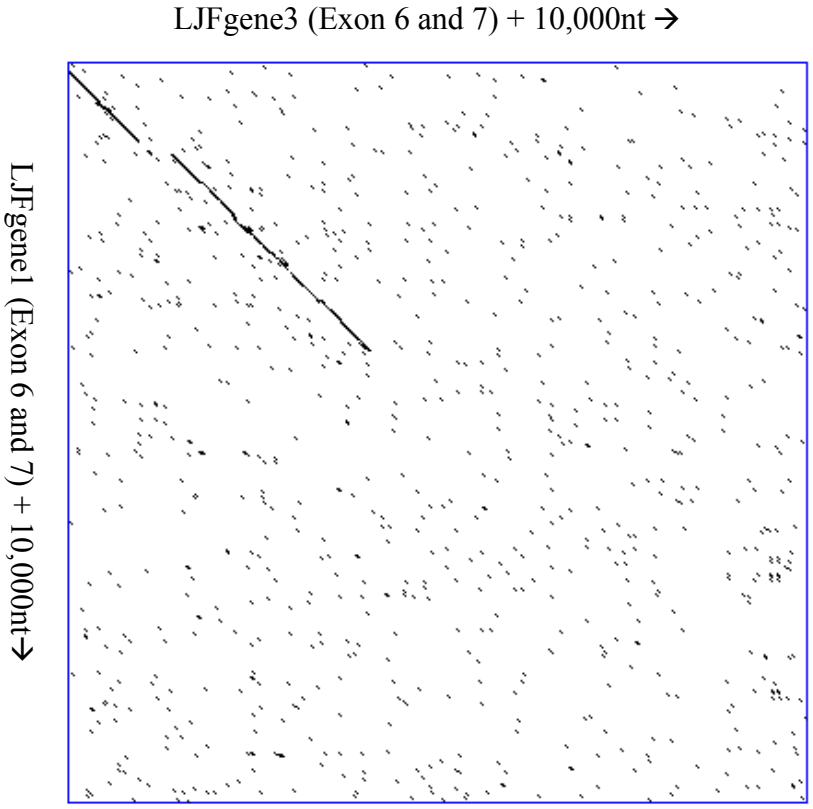


Figure 3.14. Dot plot: LJFgene3 vs. LJFgene1. Adjustment: approximately 10 kbp sequence from chromosomes 3 (x-axis) and chromosome 1 (y-axis) beginning at the second to last exon of each gene family model and extending past the 3' end.

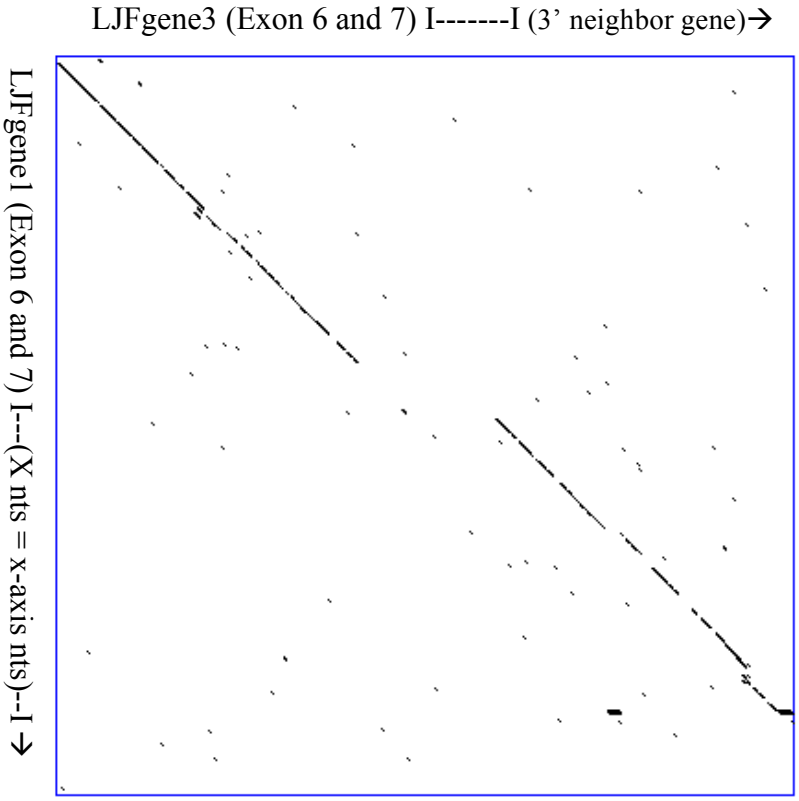


Figure 3.15. Dot plot: LJFgene3 vs. LJFgene1. Adjustment: sequence from chromosomes 3 (x-axis) and chromosome 1 (y-axis) beginning at the second to last exon of each gene family model and extending past 3' gene model boundary up to immediately preceding nearest 3' neighbor gene on chromosome 3.

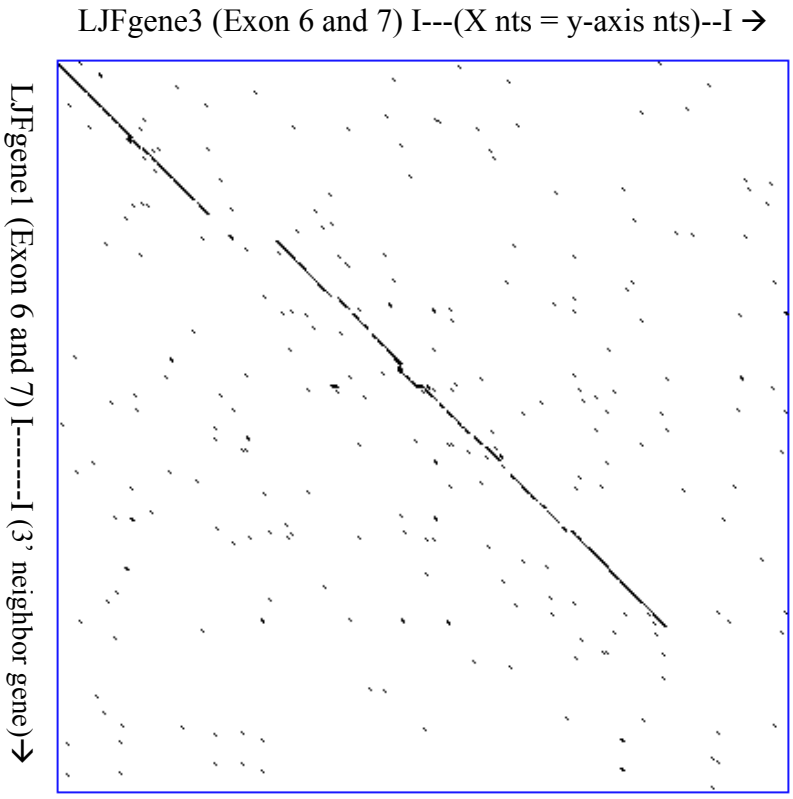


Figure 3.16. Dot plot: LJFgene3 vs. LJFgene1. Adjustment: sequence from chromosomes 3 (x-axis) and chromosome 1 (y-axis) beginning at the second to last exon of each gene family model and extending past 3' gene model boundary up to immediately preceding nearest 3' neighbor gene on chromosome 1.

3.4. FUNCTIONAL ANALYSIS

The effort to identify the function of the putative protein for the LJFgene family involves analysis of sequence for conserved motifs, analysis of known DNA elements associated with transcription, subcellular localization predictions, and threading sequence against databases of known proteins to predict structure.

3.4.1. Domain Identification Through Conservation of Sequence. Figure 3.17 shows the results of a conserved motif analysis for the LJFgene family.

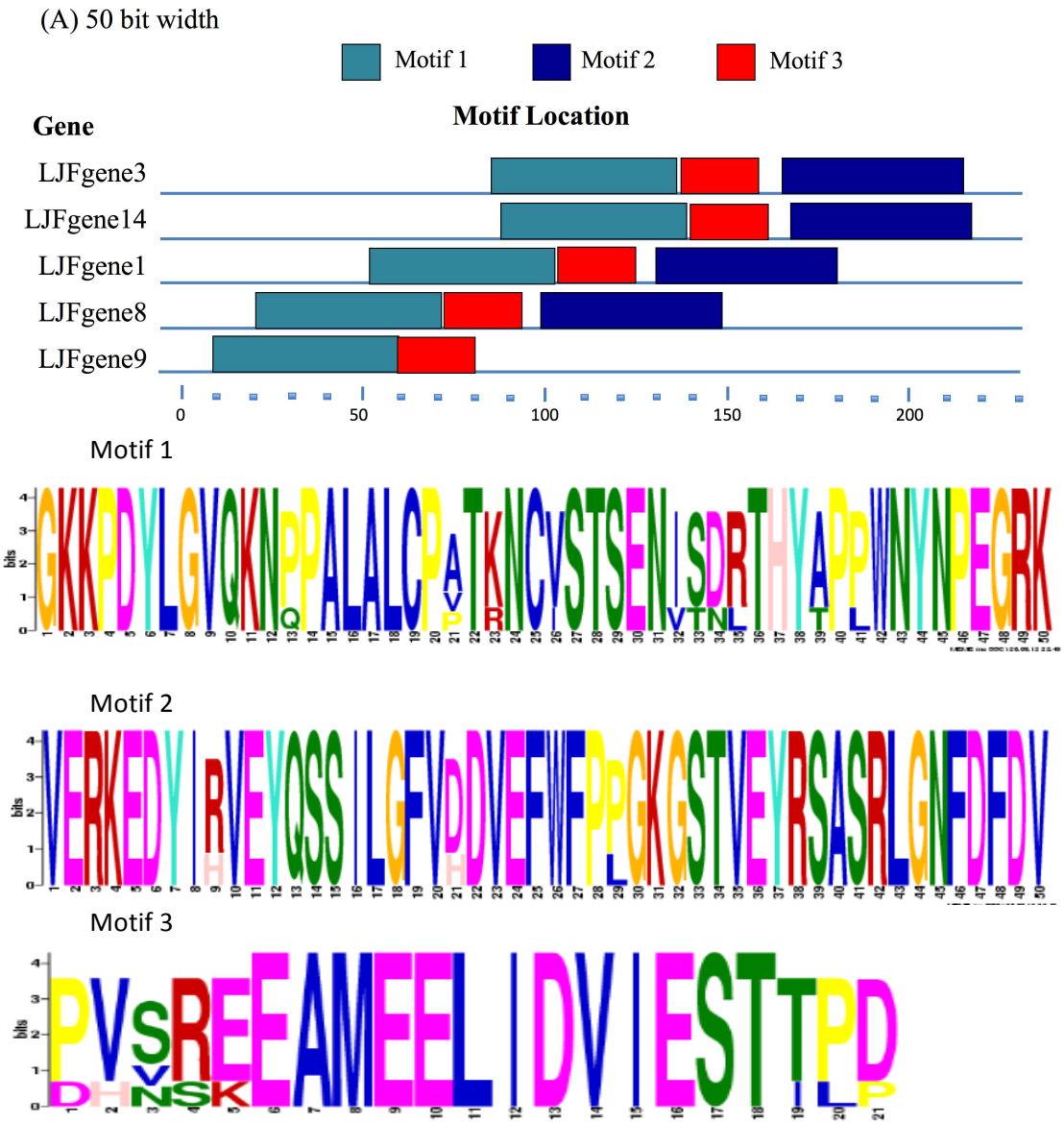


Figure 3.17. LJFgene family conserved motifs search results. (A) Motif search using 50 bit parameter (maximum length of motif 50 residues). Block diagram illustrating position of motifs along peptide length accompanied by proportional amino acid composition at each position of the motif. (B) Motif search using 100 bit parameter (maximum length of motif 100 residues). Block diagram illustrating position of motifs along peptide length accompanied by proportional amino acid composition at each position of the motif.

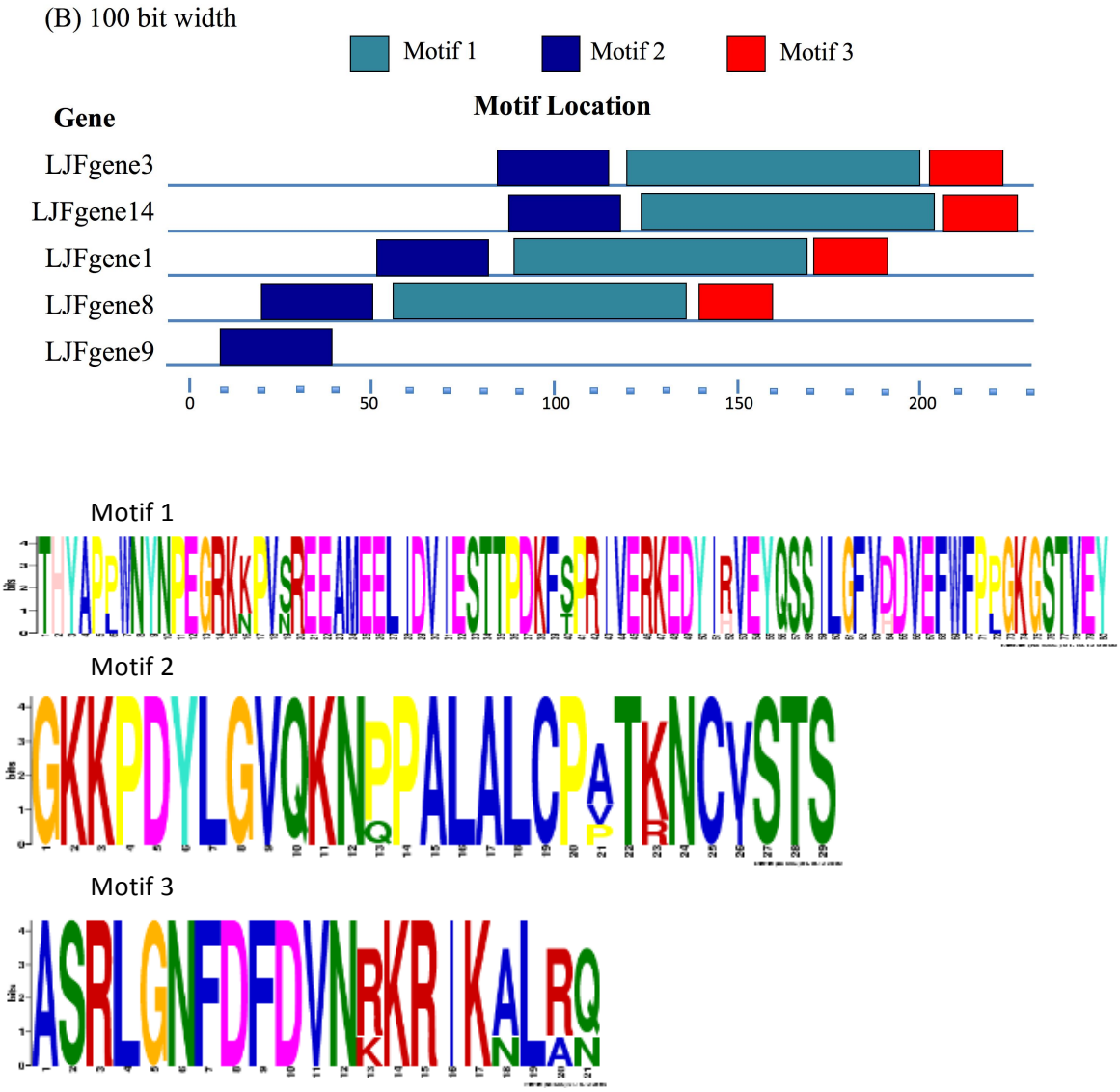


Figure 3.17. LJFgene family conserved motifs search results. (CONT.)

3.4.2. Promoter Element Analysis. The cis-element data has been organized in two forms. Those elements that fall into one of three categories according to gene association (only associated with genes with EST data, only associated with genes without EST data, or associated with all family members) are organized in Table 3.8. Cis-elements that have tissue-specific or treatment-specific themes with overlapping appearance in gene family sequences are organized into Table 3.10.

Table 3.8. Plant cis-acting elements upstream of LJFgene family members.

Identifier	Element	Genes w/ ESTs			Genes w/o ESTs	
		LJFgene3	LJFgnee14	LJFgene9	LJFgene1	LJFgene8
S000353	AACAAAC	✓	✓	✓		
S000414	ACGTG	✓	✓	✓		
S000507	MACGYGB	✓	✓	✓		
S000499	GAGAC	✓	✓	✓		
S000458	AACGTG	✓	✓	✓		
S000400	TATTTAA	✓	✓	✓		
S000308	AAACCCTAA				✓	✓
S000472	AAACCCTA				✓	✓
S000459	GNATATNC				✓	✓
S000415	ACGT	✓	✓	✓	✓	✓
S000454	NGATT	✓	✓	✓	✓	✓
S000028	CAAT	✓	✓	✓	✓	✓
S000449	YACT	✓	✓	✓	✓	✓
S000030	CCAAT	✓	✓	✓	✓	✓
S000493	GTAC	✓	✓	✓	✓	✓
S000265	AAAG	✓	✓	✓	✓	✓
S000292	ACACNNG	✓	✓	✓	✓	✓
S000144	CANNTG	✓	✓	✓	✓	✓
S000494	GANTTNC	✓	✓	✓	✓	✓
S000039	GATA	✓	✓	✓	✓	✓
S000198	GRWAAW	✓	✓	✓	✓	✓
S000453	GAAAAA	✓	✓	✓	✓	✓
S000378	GTGA	✓	✓	✓	✓	✓
S000199	GATAA	✓	✓	✓	✓	✓
S000395	YTCANTYY	✓	✓	✓	✓	✓
S000067	TTWTWTTWTT	✓	✓	✓	✓	✓
S000413	CATGTG	✓	✓	✓	✓	✓
S000174	CACATG	✓	✓	✓	✓	✓
S000407	CANNTG	✓	✓	✓	✓	✓
S000462	CTCTT	✓	✓	✓	✓	✓
S000468	CTCTT	✓	✓	✓	✓	✓
S000080	AATAAA	✓	✓	✓	✓	✓
S000081	AATTAAA	✓	✓	✓	✓	✓
S000088	AATAAT	✓	✓	✓	✓	✓
S000245	AGAAA	✓	✓	✓	✓	✓
S000450	ACTCAT	✓	✓	✓	✓	✓
S000259	CCTTTT	✓	✓	✓	✓	✓
S000314	CAACA	✓	✓	✓	✓	✓
S000098	ATATT	✓	✓	✓	✓	✓
S000103	RTTTTTR	✓	✓	✓	✓	✓
S000203	TTATTT	✓	✓	✓	✓	✓
S000390	TTGAC	✓	✓	✓	✓	✓
S000442	TGACT	✓	✓	✓	✓	✓
S000457	TGACY	✓	✓	✓	✓	✓
S000447	TGAC	✓	✓	✓	✓	✓

Table 3.9 outlines treatment data from the ESTs associated with gene family members. The ESTs that were generated from cDNAs created through the sampling of specific tissues or stress induction are listed for comparison with elements in Table 3.10.

Table 3.9. Treatment data from EST library.

Gene	EST	Treatment
LJFgene3	BM176973	hypersensitive response induced with Pseudomonas
	BM886799	Drought stress treatment
	CO983876	exposure to fungal pathogens
	EV275072	Drought stressed, salt stressed and Pseudomonas-infected
LJFgene14	EV264523	apical meristem and green seeds
	FG993792	apical meristem and green seeds
	HO044862	immature seeds
LJFgene9	CF921901	root hair treated with nodulating bacteria (Bradyrhizobium)
	CF923165	root hair treated with nodulating bacteria (Bradyrhizobium)

Table 3.10. Shared and noteworthy themes of LJFgene family promoter elements.

Interesting Characteristic	Identifier	Species	LJFgene(s)
endosperm specific	S000353	*O. sativa* (rice)	3,14,9
	S000265	*P. sativum* (pea)	3,14,9,8,1
source tissue: seed	S000353	*O. sativa* (rice)	3,14,9,8,1
	S000028	*Z. mays* (corn)	3,14,9,8,1
	S000144	*B. napus* (rapeseed)	3,14,9,8,1
	S000103	*G. max* (soybean)	3,14,9,8,1
Response to dehydration	S000414	*A. thaliana* (Thale cress)	3,14,9
	S000415	*A. thaliana* (Thale cress)	3,14,9,8,1
	S000413	*A. thaliana* (Thale cress)	3,14,9,8,1
	S000174	*A. thaliana* (Thale cress)	3,14,9,8,1
	S000407	*A. thaliana* (Thale cress)	3,14,9,8,1
TATA rice PAL gene	S000400	*O. sativa* (rice)	3,14,9
Response elements	S000499	*A. thaliana* (Thale cress)	3,14,9
	S000459	*A. thaliana*, *L. esculentum*, *M. truncatula*, *H. vulgare*	8,1
	S000493	*C. reinhardtii* (green algae)	3,14,9,8,1
	S000144	*B. napus* (rapeseed)	3,14,9,8,1
	S000390	*A. thaliana* (Thale cress)	3,14,9,8,1
	S000292	*A. thaliana* (Thale cress), *D. carota* (carrot)	3,14,9,8,1
Wound-induced	S000458	*A. thaliana* (thale cress), *L. esculentum* (tomato)	3,14,9
	S000457	*N. tabacum* (tobacco)	3,14,9,8,1

Table 3.10. Shared and noteworthy themes of LJFgene family promoter elements. (cont.)

Root/ Nodules	S000308	*A. thaliana* (Thale cress)	3,14,9,8,1
	S000468	*M. truncatula* (barrel medic), *G. max* (soybean)	3,14,9,8,1
	S000098	*Agrobacterium rhizogenes*	3,14,9,8,1
Axillary bud	S000472	*A. thaliana* (Thale cress)	8,1
Light-responsive	S000039	*A. thaliana, O. sativa, P. hybrida (*petunia)	3,14,9,8,1
	S000198	*P. sativum, A. sativa, O. sativa, N. tabacum, A. thaliana, S. oleracea*, bean	3,14,9,8,1
	S000199	N/A	3,14,9,8,1
	S000395	*N. tabacum* (tobacco)	3,14,9,8,1
pathogen-induced	S000453	*G. max* (soybean)	3,14,9,8,1
	S000468	*M. truncatula* (barrel medic), *G. max* (soybean)	3,14,9,8,1
	S000447	*O. sativa* (rice), *P. crispum* (parsley)	3,14,9,8,1

3.4.3. Subcellular Localization Predictions. The location of putative protein function within the cell was predicted using multiple methods including two computer programs and a hydropathy plot.

3.4.3.1. CELLO. The results of subcellular localization predictions produced by CELLO are organized in Table 3.11 by LJFgene member and ranked according to score.

Table 3.11. CELLO results summary.

Gene Family Member	CELLO Localization rank	Localization	Reliability score
LJFgene3	1	Nuclear	2.264*
	2	Chloroplast	1.817*
	3	Mitochondrial	0.371
LJFgene14	1	Nuclear	1.171*
	2	Chloroplast	1.068*
	3	Extracellular	1.010*
LJFgene1	1	Nuclear	2.391*
	2	Chloroplast	1.125
	3	Mitochondrial	0.633
LJFgene8	1	Nuclear	1.877*
	2	Mitochondrial	0.967
	3	Plasma Membrane	0.794
LJFgene9	1	Nuclear	1.276*
	2	Cytoplasmic	1.065*
	3	Extracellular	0.894

*designates significant scores indicated by CELLO output.

3.4.3.2. Hydropathicity analysis. Hydropathy plots were used to assess whether the conceptually translated amino acid sequence of LJFgene3 would fit the criteria for an integral membrane protein. The hydropathy plot of LJFgene3, displayed in Figure 3.18, was compared to the hydropathy plot of a known integral membrane protein, human rhodopsin, which is displayed in Figure 3.19.

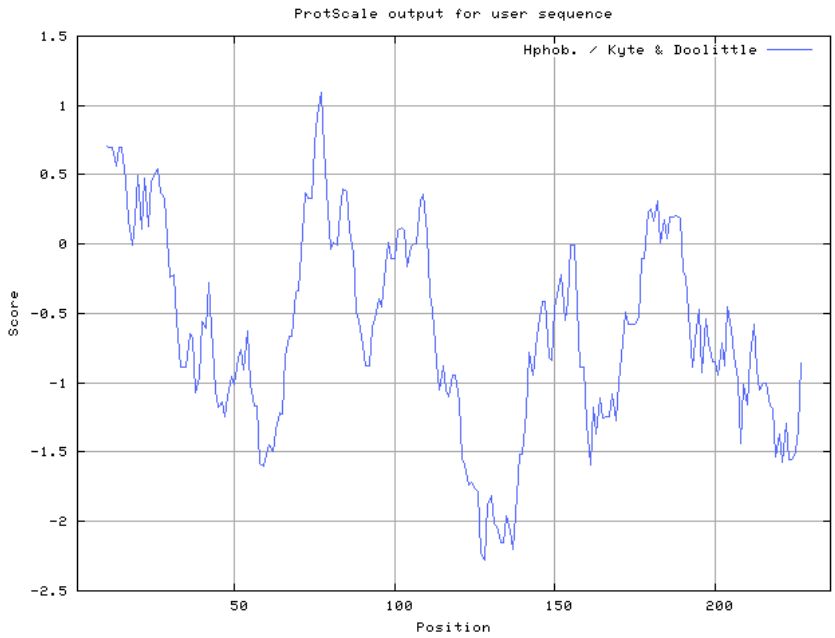


Figure 3.18. Kyte-Doolittle hydropathy plot of LJFgene3. Window size = 19. Peaks scoring >1.6 indicative of transmembrane segments.

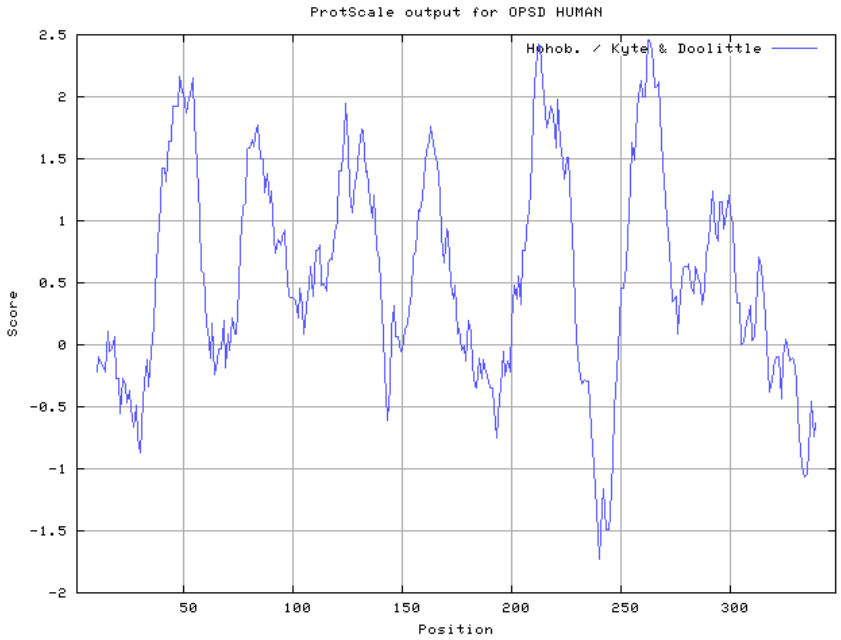


Figure 3.19. Hydropathy plot of human rhodopsin protein (known transmembrane protein). ProtScale input: Human Rhodopsin amino acid sequence accessed by UniProtKB identifier P08100 (OPSD_HUMAN) [75].

3.4.3.3. I-TASSER gene ontology results. The gene ontology data provided by I-TASSER classifies the predicted protein product of LJFgene3 as follows:

- Ontology: Cellular Component
- GO:0071944 Cell Periphery
- Definition: The part of the cell encompassing the cell cortex, the plasma membrane, and any external encapsulating structures.

3.4.4. Secondary Structure Predictions. The arrangement of an amino acid sequence into alpha helices and beta sheets was predicted using two programs. The output provides the predicted secondary structure at each loci as well as a confidence score (1 -10; higher scores indicate more confident predictions). Figure 3.20 displays the prediction according to PSIPRED. Figure 3.21 displays the prediction according to I-TASSER. A comparison of both outputs is demonstrated in Figure 3.22.

```
Conf: 961002330112235531000234554110011248999987676102201335899996
Pred: CCCCHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHCCCCCCCCCC
  AA: MSISSLIFSNLHFQLPTTMASMASSSSFCNLKFITKPNNGRRSSLPRIVFCQKHHDSTPT
              10        20        30        40        50        60

Conf: 422015789987799999999740799998877547999877999999901135556788
Pred: CCCCHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCEECCCCCCCCC
  AA: DQINRRELILRSSEIATIGAILNFGGKKPDYLGVQKNPPALALCPATKNCVSTSENISDR
              70        80        90       100       110       120

Conf: 876698746999999989989999999999985399999981784428869999970557
Pred: CCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHCCCCCCCCEEEEEECCEEEEEEEECCC
  AA: THYAPPWNYNPEGRKKPVNREEAMEELIDVIESTTPDKFSPRIVERKEDYIRVEYQSSIL
             130       140       150       160       170       180

Conf: 873108999609998469998537889995576799999999999998598556889
Pred: CCCCCEEEEECCCCCCEEEEEECCCCCCCCHHHHHHHHHHHHHHHHHCCCCCCCCC
  AA: GFVDDVEFWFPPGKGSTVEYRSASRLGNFDFDVNRKRIKALRQELEKKGWASQDTI
             190       200       210       220       230
```

Figure 3.20. Secondary structure prediction, including confidence scores at each position, of PSIPRED HFORMAT (PSIPRED V3.3) on the conceptually translated amino acid sequence of LJFgene3. Alpha helices are designated with red H's and beta sheets with blue E's.

```
Sequence   MSISSLIFSNLHFQLPTTMASMASSSSFCNLKFITKPNNGRRSS
Prediction CCHHHHHHHCCCCCCCCCHHHHHCCCCCCCCSSSSCCCCCCCCCC
Conf.Score 94667645110100463367651542302135311369733364

LPRIVFCQKHHDSTPTDQINRRELILRSSEIATIGAILNFGGKKPDYLGVQKNPP
CHHHHHCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCCC
01555353579876543534557899999999999999725899876886688 75

ALALCPATKNCVSTSENISDRTHYAPPWNYNPEGRKKPVNREEAMEELIDVIEST
CCCCCCCCCCCCSSCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHHC
66899799971746754567423458700477544455589999999999999863

TPDKFSPRIVERKEDYIRVEYQSSILGFVDDVEFWFPPGKGSTVEYRSASRLGNF
CCCCCCCSSSSCCCCSSSSSSSCCCCCCCCSSSSSSSCCCCCSSSSSSCHCCCCC
12357863662369989999971555776339999967998789999940106777

DFDVNRKRIKALRQELEKKGWASQDTI
CCCHHHHHHHHHHHHHHCCCCCCCCC
65573999999999987587575569
```

Figure 3.21. Secondary structure prediction, including confidence scores at each position, of I-TASSER on the conceptually translated amino acid sequence of LJFgene3. Alpha helices are designated with red H's and beta sheets with blue S's.

```
PSIPRED:      CCCCHHHHHCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
I-TASSER:     CCHHHHHHHCCCCCCCCCHHHHHCCCCCCCSSSSCCCCCCCCC
AA Seq:       MSISSLIFSNLHFQLPTTMASMASSSSFCNLKFITKPNNGRRSS

PSI PRED:     CCHHHHCCCCCCCCCCCCCCHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCC
I-TASSER:     CHHHHHCCCCCCCCCCHHHHHHHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCC
AA Seq:       LPRIVFCQKHHDSTPTDQINRRELILRSSEIATIGAILNFGGKKPDYLGVQKNPP

PSI PRED:     CCCCCCCCCCCCEECCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHC
I-TASSER:     CCCCCCCCCCCCSSCCCCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHHHHHHHHHHC
AA Seq:       ALALCPATKNCVSTSENISDRTHYAPPWNYNPEGRKKPVNREEAMEELIDVIEST

PSI PRED:     CCCCCCCEEEEEECCEEEEEEEECCCCCCCCEEEEECCCCCCEEEEEECCCCCCC
I-TASSER:     CCCCCCCSSSSCCCCSSSSSSSCCCCCCCCSSSSSSSCCCCCSSSSSSCHCCCCC
AA Seq:       TPDKFSPRIVERKEDYIRVEYQSSILGFVDDVEFWFPPGKGSTVEYRSASRLGNF

PSI PRED:     CHHHHHHHHHHHHHHHHCCCCCCCCC
I-TASSER:     CCCHHHHHHHHHHHHHHCCCCCCCCC
AA Seq:       DFDVNRKRIKALRQELEKKGWASQDTI
```

Figure 3.22. Alignment of prediction tool outputs to determine level of agreement. PSI PRED output designates alpha helices with red H's and beta sheets with blue E's. I-TASSER output designates alpha helices with red H's and beta sheets with Blue S's.

3.4.5. Tertiary Structure and Function Predictions. Predictions of tertiary structure and function of the LJFgene3 conceptually translated peptide sequence were conducted using numerous threading programs. The results from submission to pDomTHREADER are illustrated in Figures 3.23 and 3.24, as well as in Table 3.12. The results from submission to pGenTHREADER are illustrated in Table 3.13. The results from submission to I-TASSER are illustrated in Tables 3.14, 3.15, 3.16, 3.17, and 3.18.

(A)

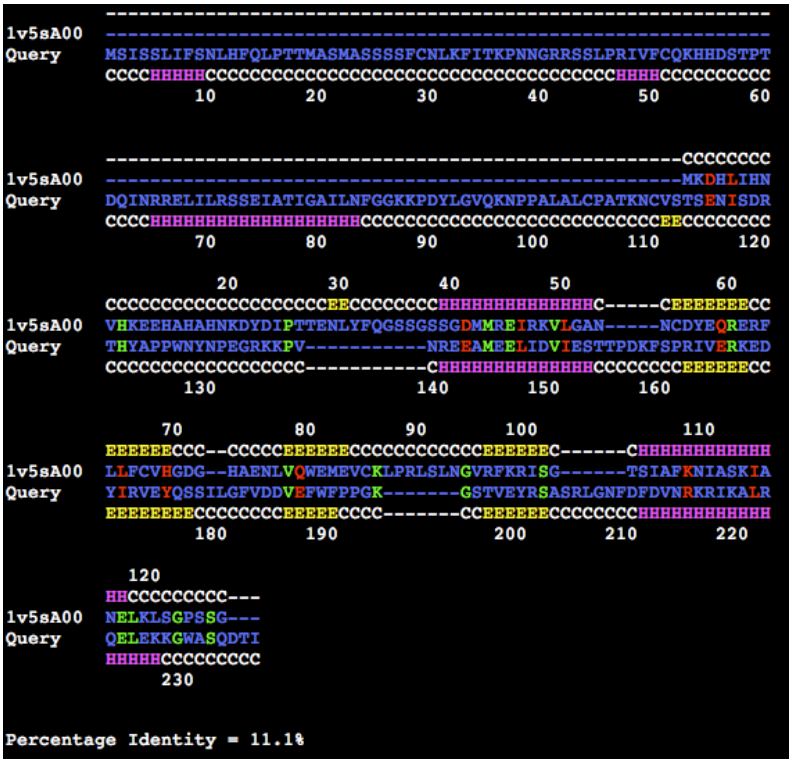


Figure 3.23. Top 3 pDomTHREADER secondary structure alignments of query sequence (LJFgene3) against domain codes based on secondary structure similarities (as opposed to alignment scores). (A) Secondary structure of the domain with code 1v5sA00 exhibits the highest level of structural similarity with the query at the carboxy- terminus of the query. (B) Secondary structure of the domain with code 1up8A00 exhibits the highest level of structural similarity with the query at the amino-terminus of the query. (C) Secondary structure of the domain with code 1m40A00 exhibits a high degree of structural similarity over the entire length of the query.

(B)

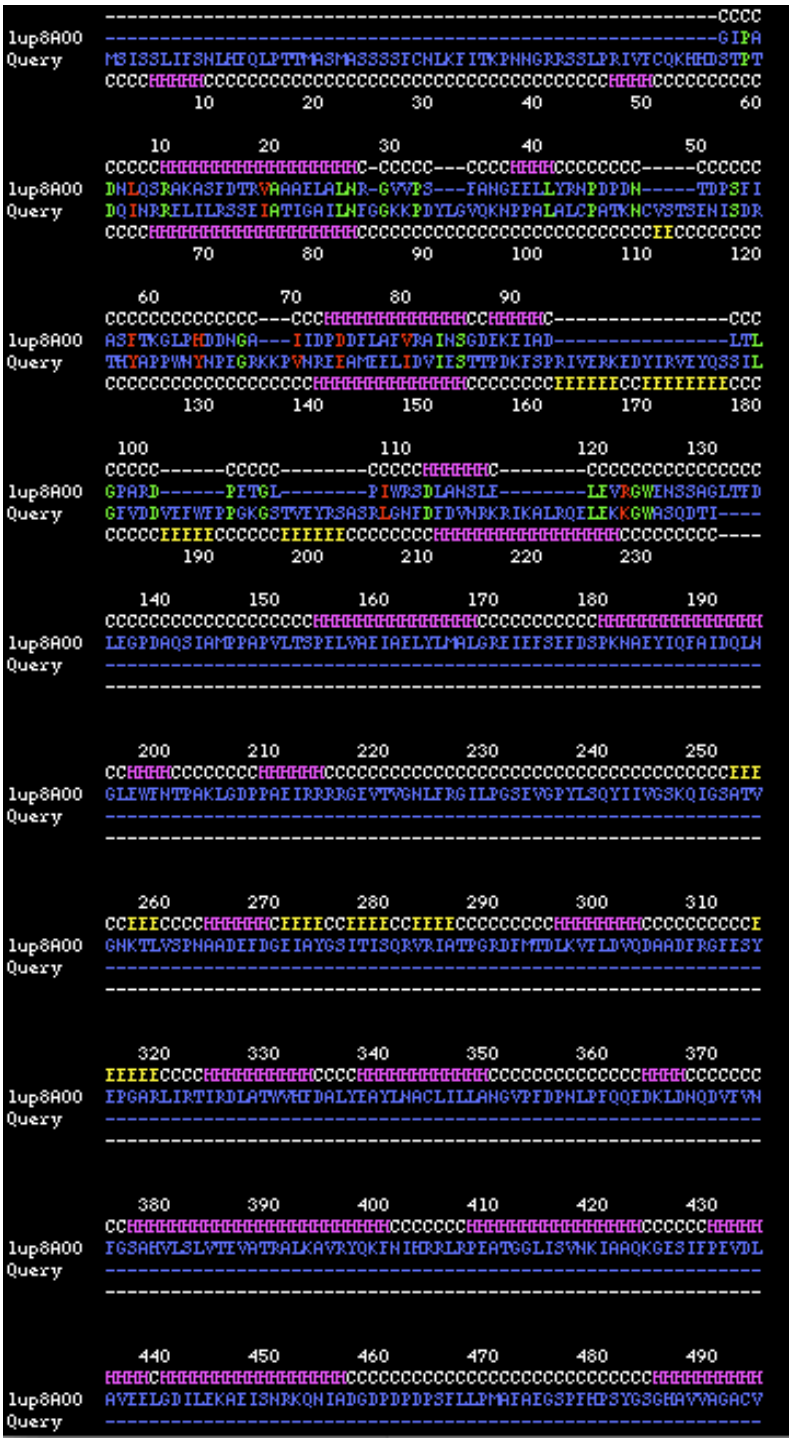


Figure 3.23. Top 3 pDomTHREADER secondary structure alignments of query sequence (LJFgene3) against domain codes based on secondary structure similarities (as opposed to alignment scores). (CONT.)

(C)

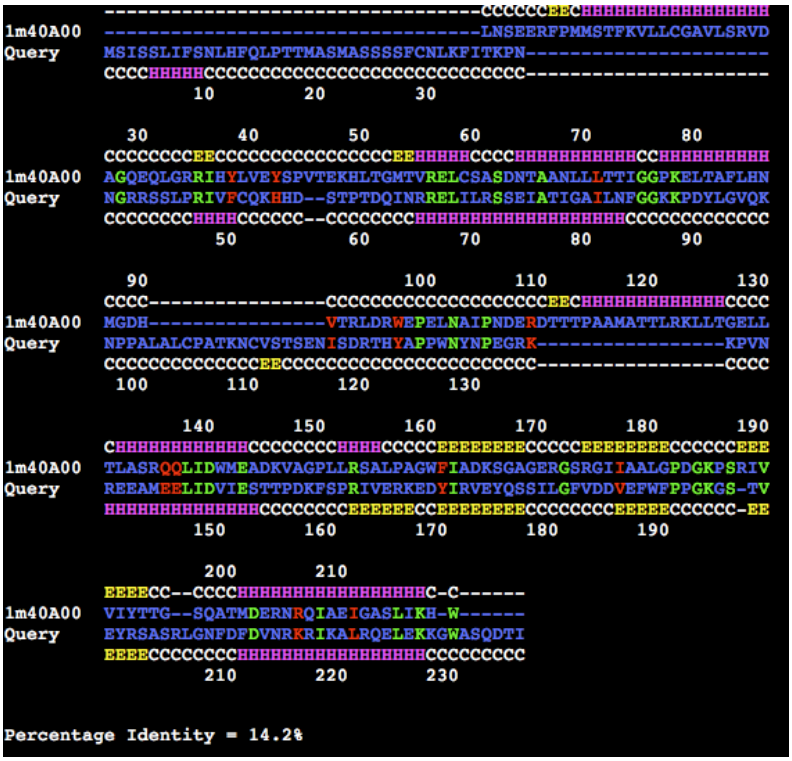


Figure 3.23. Top 3 pDomTHREADER secondary structure alignments of query sequence (LJFgene3) against domain codes based on secondary structure similarities (as opposed to alignment scores). (CONT.)

Table 3.12. CATH domain summary of pDomTHREADER output.

CATH Level	Description	Qty
Class	Alpha Beta	8
	Mainly Beta	1
	Mainly Alpha	1
Architecture	Roll	3
	2-layer Sandwich	2
	3-layer (aba) Sandwich	2
	Sandwich	1
	Orthogonal Bundle	1
	Alpha-Beta Complex	1
Topology	No Agreement	
Homologous Superfamily	No Agreement	

(A) 1v5sA00	**Level**	**CATH Code**	**Description**
	C	3	Alpha Beta
	A	3.30	2-Layer Sandwich
	T	3.30.310	TATA-binding Protein
	H	3.30.310.80	Kinase-associated Domain

(B) 1up8A00	**Level**	**CATH Code**	**Description**
	C	1	Mainly Alpha
	A	1.10	Orthogonal Bundle
	T	1.10.606	Vanadium-containing Chloroperoxidase
	H	1.10.606.10	Vanadium-containing Chloroperoxidase

(C) **1m40A00**	**Level**	**CATH Code**	**Description**
	C	3	Alpha Beta
	A	3.40	3-layer (aba) Sandwich
	T	3.40.710	Beta-lactamase
	H	3.40.710.10	DD-peptidase/ β-lactamase

Figure 3.24. CATH classification for the 3 pDomTHREADER domains with the most secondary structure similarity. (A) 1v5sA00 [76], (B) 1up8A00 [77], and (C) 1m40A00 [78].

Table 3.13. Summary of pGenTHREADER results.

PDB identifier	Identification Method	Host Organism	Molecular Classification	Molecule
2Y94/ 4CFH	X-ray Diffraction	*Escherichia coli*	Transferase	5'-AMP-activated protein kinase catalytic subunit
2EBM	Solution NMR	Not Listed	Unknown Function	RWD domain containing protein
2RRL	Solution NMR	*Escherichia coli*	Protein Transport	Flagellar hook-length control protein
2FSQ	X-ray Diffraction	*Escherichia coli* BL21	Unknown Function	Putative uncharacterized protein
3TOD	X-ray Diffraction	Not Listed	Hydrolase	C-lobe of bovine lactoferin
2JOI	Solution NMR	*Escherichia coli*	Unknown Function	Putative uncharacterized protein
2VZ8	X-ray Diffraction	Not Listed	Transferase	Fatty acid synthase
4FR9	X-ray Diffraction	*Escherichia coli*	Unknown Function	Putative uncharacterized protein
3OAJ	X-ray Diffraction	*Escherichia coli*	Unknown Function	Putative dioxygenase
1DOT	X-ray Diffraction	Not Listed	Iron Transport	ovotransferrin

Table 3.14. Summary of I-TASSER results: Top 10 threading templates.

Rank	PDB Identifier	Molecular Classification	Molecule	Identity
1	3w4qA	Hydrolase	Beta-lactamase	0.14
2	3w4qA	Hydrolase	Beta-lactamase	0.09
3	3w4qA	Hydrolase	Beta-lactamase	0.13
4	4gbmA	Transferase	Sulfotransferase	0.12
5	3mekA	Transferase	Methyltransferase	0.11
6	4btgA	Viral protein	Capsid coordination	0.13
7	4m5uA	RNA-binding/inhibitor	Polymerase PA	0.14
8	2kixA	Transport protein	BM2 protein	0.46
9	3fleA	Unknown function	Unknown function	0.08
10	1ef1C	Membrane protein	Moesin	0.15

* Identity is the percentage of similarity between query and the aligned region of the templates.

Table 3.15. Summary of I-TASSER results: Top 10 structural analogs.

Rank	PDB Identifier	Molecular Classification	Molecule	Identity	TM-score
1	3w4qA	Hydrolase	Beta-lactamase	0.129	0.876
2	3bydA	Hydrolase	Beta-lactamase OXY-1	0.121	0.871
3	3w4oA	Hydrolase	Beta-lactamase	0.098	0.870
4	1hzoA	Hydrolase	Beta-lactamase	0.115	0.870
5	1bsg	Hydrolase	Beta-lactamase	0.111	0.870
6	3dw0B	Hydrolase	Class A Beta-lactamase KPC-2	0.122	0.869
7	4eqiA	Hydrolase	Class A Beta-lactamase SFC-1	0.115	0.869
8	1iyqA	Hydrolase	Toho-1 Beta-lactamase	0.117	0.868
9	1dy6B	Hydrolase	Class A Beta-lactamase SME-1	0.107	0.864
10	1bueA	Hydrolase	Imipenem-hydrolysing Beta-lactamase	0.103	0.864

* Rank is based on TM-scores.
* TM-score measures structural similarity between template and query.
* Identity is the percentage of sequence similarity within structurally aligned regions.

Table 3.16. Summary of I-TASSER results: Top 5 enzyme homologs.

Rank	PDB ID	$Cscore^{EC}$	TM-score	EC #	Protein Classification	Molecule
1	1iysA	0.215	0.868	3.5.2.6	Hydrolase	Beta-lactamase
2	3lezA	0.209	0.863	3.5.2.6	Hydrolase	Beta-lactamase
3	3c4pA	0.189	0.845	3.5.2.6	Hydrolase	Beta-lactamase
4	3bydA	0.176	0.871	3.5.2.6	Hydrolase	Beta-lactamase
5	1iyqA	0.175	0.868	3.5.2.6	Hydrolase	Beta-lactamase

* $Cscore^{EC}$ is a measure of confidence in the EC number prediction. Scores range from 0 to 1, with numbers closer to 1 indicating more reliable predictions.
* TM-score measures structural similarity between template and query.

Table 3.17. Summary of I-TASSER results: gene ontology prediction.

	GO Term	GO score	Description
Molecular Function	GO:0008800	0.71	Beta-lactamase activity
Biological Process	GO:0046677	0.71	Response to antibiotic
	GO:0030655	0.71	Beta-lactam antibiotic catabolic process
Cellular Location	GO:0071944	0.41	Cell periphery

* GO score is assigned based on weighted $C\text{score}^{GO}$ scores for the GO terms. Scores range from 0 to 1, with numbers closer to 1 indicating more reliable predictions.

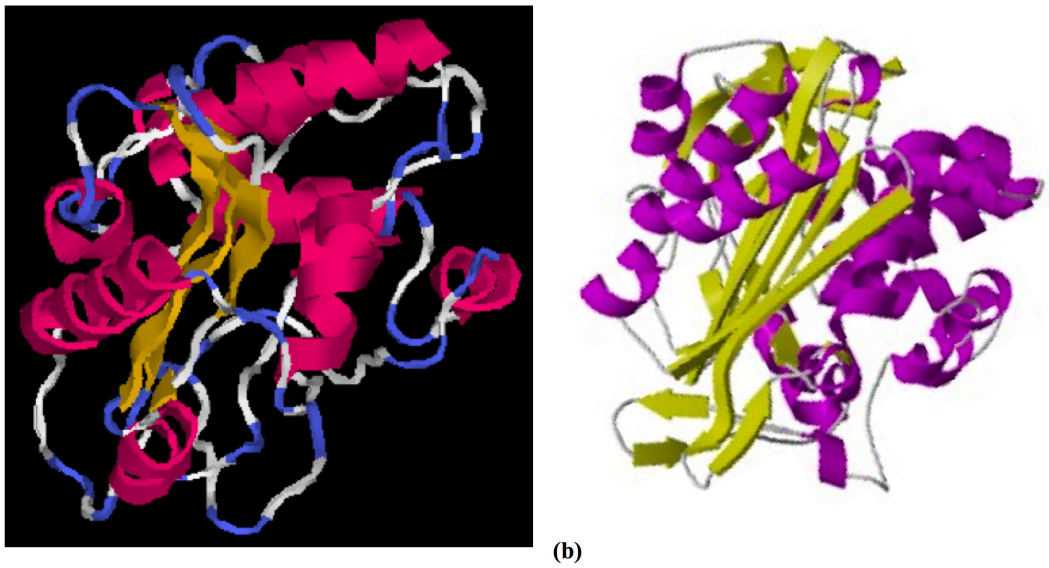
Table 3.18. Summary of I-TASSER results: Top 10 templates with binding sites similar to the query.

Rank	PDB ID	$C\text{score}^{LB}$	BS-score	Predicted BS Residues	Protein Classification	Molecule
1	3sh8B	0.50	0.93	72, 74, 175 -178, 183	Hydrolase/ antibiotic	Beta-lactamase
2	3hlwA	0.46	1.02	72, 74, 108, 113, 156, 174, 175 - 179	Hydrolase	Beta-lactamase
3	3b3xB	0.27	0.96	72, 175, 177, 183, 213	Hydrolase/ inhibitor	Beta-lactamase
4	3m6hA	0.24	0.94	20, 49, 52, 72, 107, 175 - 177	Hydrolase/ antibiotic	Beta-lactamase
5	1blcA	0.23	0.94	72, 156, 174 – 177	Hydrolase	Beta-lactamase
6	3ny4A	0.04	1.13	205, 210 – 212, 214	Hydrolase/ antibiotic	Beta-lactamase
7	1jtd0	0.04	1.02	38, 39, 41, 52, 53, 55, 71, 156, 177	Hydrolase/ inhibitor	Beta-lactamase
8	1jtg0	0.04	1.00	38, 39, 41, 52 - 54, 71, 72, 108, 113, 156, 174 – 179, 183	Hydrolase	Beta-lactamase
9	3g35B	0.04	0.96	84, 85, 90, 91	Hydrolase/ inhibitor	Beta-lactamase
10	3huoB	0.04	1.05	144, 145, 148, 149, 152, 190	Hydrolase	Beta-lactamase

*$C\text{score}^{LB}$ is a measure of confidence in the prediction of the binding site. Scores range from 0 to 1, with numbers closer to 1 indicating more reliable predictions.
* BS-score is a measure of structural and sequence similarity between query and template binding sites. Scores >1 are considered a significant local match.

The 3D protein model predicted for LJFgene3 by I-TASSER is displayed in Figure 3.25. A comparison of this structure with the structure of an experimentally classified class C beta-lactamase molecule is illustrated in Figure 3.26.

Figure 3.25. Top I-TASSER generated model for LJFgene3. Confidence score = -2.72. C-score range is [-5, 2]; where 2 is the highest confidence and -5 the lowest.



(a) (b)

Figure 3.26. Side-by-side comparison of tertiary structure of LJFgene3 predicted model and beta-lactamase molecule. (a) LJFgene3 model prediction generated by I-TASSER. (b) Class C beta-lactamase molecule from *Enterobacter cloacae* experimentally characterized by x-ray diffraction (PDB identifier 1ga0A00) [79].

3.5. NON-CODING SEQUENCE ANALYSIS

3.5.1. Nucleotide Sequences, Amino Acid Translations, and Putative Models for Non-coding Sequences Associated with LJFgene Family.

The sequence data for non-coding sequences LJFnm19, LJFnm11, and LJFnm12 are organized in Table 3.19. The non-coding sequences were not predicted as genes by FgenesH or GenomeScan algorithms; however, they are represented by models in Figure 3.27 for comparison with LJFgene3. Figure 3.28 contains alignments of the LJFnm sequences (amino acid, coding, and genomic sequences) as an illustration of sequence conservation maintained both within possible coding regions and within intronic regions, as well as between the translated gene product.

Table 3.19. LJFnm sequences.

Name	Seq. Type	Sequences
LJFnm19	Amino acid	QFWFPPGKGSTVEYRSASRLGNFDFDVNRKRIKALRQELEKKGWTSQDTI
	Coding	TTCAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTC GGTTGGGAAACTTTGATTTTGATGTGAATAGAAAAAGAATAAAGGCATTGAGACAAG AGTTGGAGAAGAAAGGATGGACATCTCAAGATACCATATGA
	Genomic	TTCAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTC GGTTGGGAAACTTTGATTTTGATGTGAATAGAAAAAGAATAAAGGTATGATTTCATA ATTAGTATGTGCTTTCTCTATAGTTAGAATAAAGGTACTCCCCTTCCTTCATGTCAT GTCAAACATTTTATACTTAAGTAAATTCACTAAATTTTAGTCTCAAATGTTTTAACT TTATTCTAAATTAGTCACTTATTTTAACTGAAGGTAAATTTGGTTGACTATGATCAG AAATACATTGATATTTTTTAATTGGTAGAGATAAAGAATATTTTTTATGTACAATAA AGAGAGTATTTACTCCAGAGGATGCAAATCCCTTACTAAATATTTTTGTGATGAAAA ATCTTGGTTGCTGACAGGCATTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTC AAGATACCATATGATTAATAAACTCAGGCAGAATTAACATCAACATCTAAGCAAATA TTATTTCATATACTTTGTGACCTTGTATACTTTTGTATTAGATACAAATCACACAGG ATCATTTCAAGCAAACTTTTCTTAGATTTTAGGAATTGTAGAGAAATCATTGAGACA ATACTTTAAACTCTCGGGGAAGGAATGGAATGAAGACCTTG
LJFnm11	Amino acid	QFWFPPGKGSTVEYRFASRLGNFDFDVNRKRIKALRQELEKKGWTSQDTI
	Coding	TTCAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGGAGTATCGATTTGCATCTC GGTTGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGCACTGAGACAAG AGTTGGAGAAGAAAGGATGGACATCTCAAGATACCATATGA
	Genomic	TTCAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGGAGTATCGATTTGCATCTC GGTTGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATAATTTCATA ATTAATATGTGCTTTCTTTATAGTTAGATAAAGAAATTCTTGGTTCCAGGGTTATAC TCCCCTTCCTTCATGTCATGTCAAACATTTTATACTTAAGTAGATTCACTAAATTTG AGTCTGAAATGTTTTAACTTTATTCTAAATTAGTCACTTATTTTAACTGAAGGTAAA TTTGGTTGACTATGATCAGAAATACACTGATATTTTTTAATTGGTAGAGATAAAGAA TATTTTTTATGTACAATAAAGAGAGTATTTACTCCAGAGGATGCAAATCCTTTACTA AATATTTTTGTGATTAAAAATCTTGGTTGCTGACAGGCACTGAGACAAGAGTTAGAG AAGAAAGGATGGACATCTCAAGATACCATATGATTAATAAACTCAGGCAGAATTAAC ATCAGCATCTAAGCAAATATTATTTCATATACTTTGTGACCTTGTATACTTTTGTAT TAGATACAAATCGCACAGGATCATTGCAAGCAAACTTTTCTTAGATTTTTGGAATTG TAGAGAAATCATTGAGAACAATACTTCAAACTCTCGGGGAAGGAATGAAATGAAGAC CTTG
LJFnm12	Amino acid	FWFPPGKGSTVKYRSASRLGNFDFDVNRKRIKALRQELEKKGWTSQDTI
	Coding	TTTAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGAAGTATCGATCTGCATCTC GGTTGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGCACTGAGACAAG AGTTGGAGAAGAAAGGATGGACATCTCAAGATACCATATGA
	Genomic	TTTAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGAAGTATCGATCTGCATCTC GGTTGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATGATTTCATA ATTAATATGTGCTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCTAGGGTTATAC TCCCCTTCCCTCATGTGATGTCAAACATTTTATACTTAAGTAGATTCACTAAATTTG AGTCTCAAATGTTTTAACATTATTCTAAATTAGTCACTTATTTTAACTGAAGGTAAA TTTGGTTGACTATGATCAGAAATACATTGATATTTTTTAATTGGTAGAGATAAAGAA TATTTTTGATGTACAATAAAGAGAGTACTTACTCCAGAGGATGCAAATCCCTTACTA AATATTTTTGTGATGAAAAATCTTGGTTGCTGAAAGGCACTGAGACAAGAGTTGGAG AAGAAAGGATGGACATCTCAAGATACCATATGATTAATAAACTCAGGCAGAATTAAC ATCAGCATCTAAGCAAATATTATTTCATATACTTTGTGACCTTGTATACTTTTGTAT TAGATACAAATTGCAAGCAAACTTTTCTTAGATTTTTG

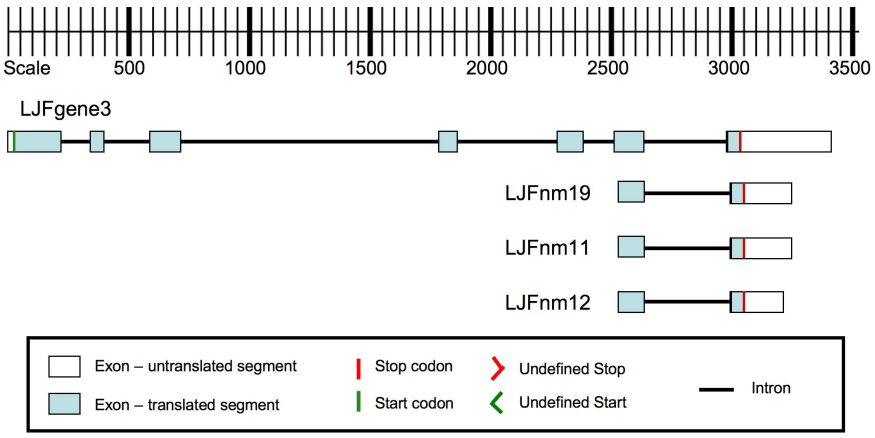


Figure 3.27. LJFnm gene models.

(A)

```
LJFnm19          QFWFPPGKGSTVEYRSASRLGNFDFDVNRKRIKALRQELEKKGWTSQDTI
LJFnm11          QFWFPPGKGSTVEYRFASRLGNFDFDVNRKRIKALRQELEKKGWTSQDTI
LJFnm12          -FWFPPGKGSTVKYRSASRLGNFDFDVNRKRIKALRQELEKKGWTSQDTI
                  ***********:** *********************************
```

(B)

```
LJFnm11          TTCAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGGAGTATCGATTTGCATCTCGGT
LJFnm19          TTCAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTCGGT
LJFnm12          TTTAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGAAGTATCGATCTGCATCTCGGT
                 ** ********************************** ********* ***********

LJFnm11          TGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGCACTGAGACAAGAGTTGG
LJFnm19          TGGGAAACTTTGATTTTGATGTGAATAGAAAAAGAATAAAGGCATTGAGACAAGAGTTGG
LJFnm12          TGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGCACTGAGACAAGAGTTGG
                 ************************ ***************** ***************

LJFnm11          AGAAGAAAGGATGGACATCTCAAGATACCATATGA
LJFnm19          AGAAGAAAGGATGGACATCTCAAGATACCATATGA
LJFnm12          AGAAGAAAGGATGGACATCTCAAGATACCATATGA
                 **********************************
```

Figure 3.28. LJFnm sequence alignments. (A) Alignment of LJFnm conceptually translated peptide sequences. (B) Alignment of LJFnm coding sequences. (C) Alignment of LJFnm genomic sequences. A star (*) indicates 100 percent identity at a loci; a colon (:) represents strong chemical property conservation between residues at a position (based on a scoring matrix threshold).

(C)

```
LJFnm11         TTCAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGGAGTATCGATTTGCATCTCGGT
LJFnm12         TTTAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGAAGTATCGATCTGCATCTCGGT
LJFnm19         TTCAGTTCTGGTTTCCACCGGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTCGGT
                ** ********************************** ********* ***********

LJFnm11         TGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATAATTTCATAATTAAT
LJFnm12         TGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATGATTTCATAATTAAT
LJFnm19         TGGGAAACTTTGATTTTGATGTGAATAGAAAAAGAATAAAGGTATGATTTCATAATTAGT
                ************************ ****************** ************ *

LJFnm11         ATGTGCTTTCTTTATAGTTAGATAAAGAAATTCTTGGTTCCAGGGTTATACTCCCCTTCC
LJFnm12         ATGTGCTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCTAGGGTTATACTCCCCTTCC
LJFnm19         ATGTGCTTTCTCTATAGTTAGAATAAAGG-------------------TACTCCCCTTCC
                *********** **********  **                      ************

LJFnm11         TTCATGTCATGTCAAACATTTTATACTTAAGTAGATTCACTAAATTTGAGTCTGAAATGT
LJFnm12         CTCATGTGATGTCAAACATTTTATACTTAAGTAGATTCACTAAATTTGAGTCTCAAATGT
LJFnm19         TTCATGTCATGTCAAACATTTTATACTTAAGTAAATTCACTAAATTTTAGTCTCAAATGT
                 ****** ************************ ************* ***** ******

LJFnm11         TTTAACTTTATTCTAAATTAGTCACTTATTTTAACTGAAGGTAAATTTGGTTGACTATGA
LJFnm12         TTTAACATTATTCTAAATTAGTCACTTATTTTAACTGAAGGTAAATTTGGTTGACTATGA
LJFnm19         TTTAACTTTATTCTAAATTAGTCACTTATTTTAACTGAAGGTAAATTTGGTTGACTATGA
                ****** **************************************************

LJFnm11         TCAGAAATACACTGATATTTTTTAATTGGTAGAGATAAAGAATATTTTTTATGTACAATA
LJFnm12         TCAGAAATACATTGATATTTTTTAATTGGTAGAGATAAAGAATATTTTTGATGTACAATA
LJFnm19         TCAGAAATACATTGATATTTTTTAATTGGTAGAGATAAAGAATATTTTTTATGTACAATA
                *********** ********************************** **********

LJFnm11         AAGAGAGTATTTACTCCAGAGGATGCAAATCCTTTACTAAATATTTTTGTGATTAAAAAT
LJFnm12         AAGAGAGTACTTACTCCAGAGGATGCAAATCCCTTACTAAATATTTTTGTGATGAAAAAT
LJFnm19         AAGAGAGTATTTACTCCAGAGGATGCAAATCCCTTACTAAATATTTTTGTGATGAAAAAT
                ********* ********************** ******************* ******

LJFnm11         CTTGGTTGCTGACAGGCACTGAGACAAGAGTTAGAGAAGAAAGGATGGACATCTCAAGAT
LJFnm12         CTTGGTTGCTGAAAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTCAAGAT
LJFnm19         CTTGGTTGCTGACAGGCATTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTCAAGAT
                ************ ***** ************* *************************

LJFnm11         ACCATATGATTAATAAACTCAGGCAGAATTAACATCAGCATCTAAGCAAATATTATTTCA
LJFnm12         ACCATATGATTAATAAACTCAGGCAGAATTAACATCAGCATCTAAGCAAATATTATTTCA
LJFnm19         ACCATATGATTAATAAACTCAGGCAGAATTAACATCAACATCTAAGCAAATATTATTTCA
                ************************************ **********************

LJFnm11         TATACTTTGTGACCTTGTATACTTTTGTATTAGATACAAATCGCACAGGATCATTGCAAG
LJFnm12         TATACTTTGTGACCTTGTATACTTTTGTATTAGATACAAAT-------------TGCAAG
LJFnm19         TATACTTTGTGACCTTGTATACTTTTGTATTAGATACAAATCACACAGGATCATTTCAAG
                ****************************************             * ****

LJFnm11         CAAACTTTTCTTAGATTTTTGGAATTGTAGAGAAATCATTGAGAACAATACTTCAAACTC
LJFnm12         CAAACTTTTCTTAGATTTTTG-------------------------------------
LJFnm19         CAAACTTTTCTTAGATTTTAGGAATTGTAGAGAAATCATTGAGACAATACTTTAAACTCT
                ****************** *

LJFnm11         TCGGGGAAGGAATGAAATGAAGACCTTG
LJFnm12         ----------------------------
LJFnm19         CGGGGAAGGAATGGAATGAAGACCTTG-
```

Figure 3.28. LJFnm sequence alignments. (CONT.)

3.5.2. Motif Conservation. When submitted along with confirmed LJFgene family members to a conserved motif identification program, the LJFnm sequences all correspond to a single motif identified in four out of five of the LJFgenes. The motif location and sequence is illustrated in Figure 3.29.

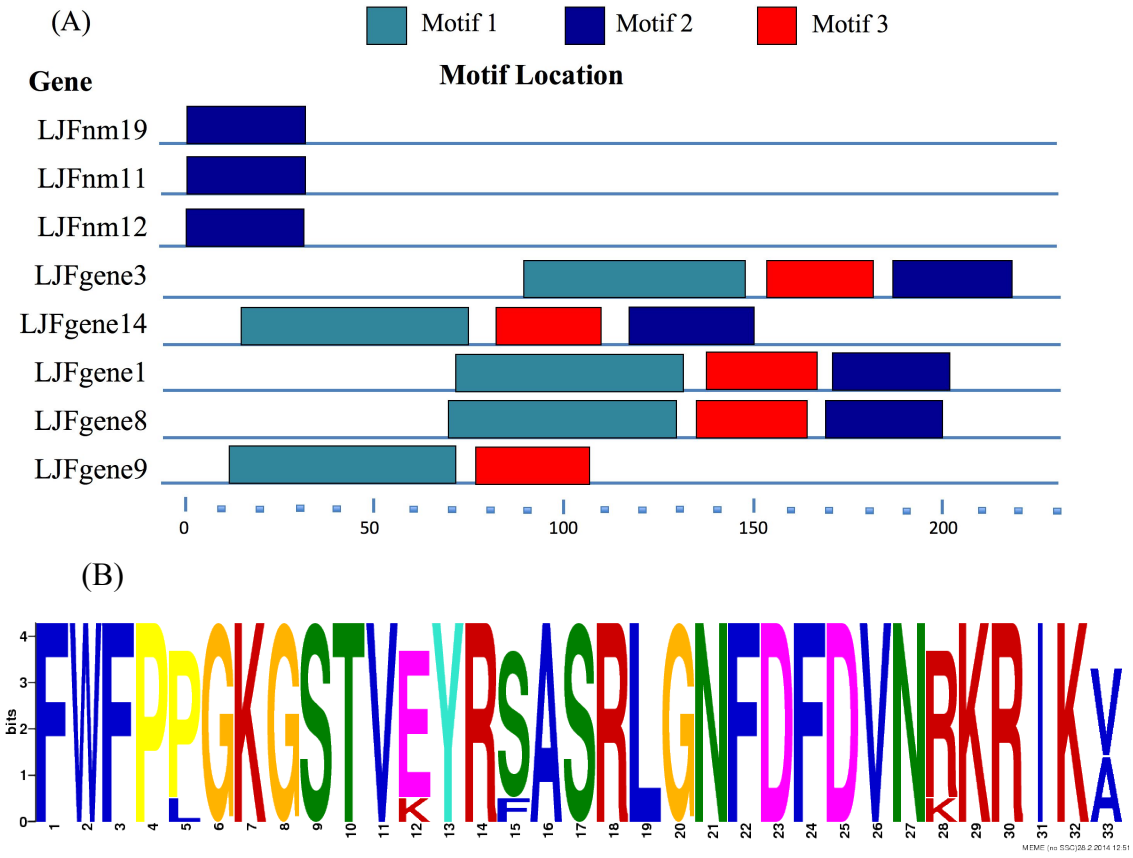


Figure 3.29. LJFnm motif search results. (A) Block diagram illustrating 100 bit width search for conserved motifs of LJFgene family as well as associated non-coding sequences. (B) Motif 2.

3.5.3. Alignment and Dot Plot of LJFnm's Against LJFgene(s). Figure 3.30 contains the output from the alignment of LJFgene3, LJFgene14, and LJFgene1 (last

two exons, last intron, and sequence extending past STOP codon) against LJFnm19, LJFnm11, and LJFnm12.

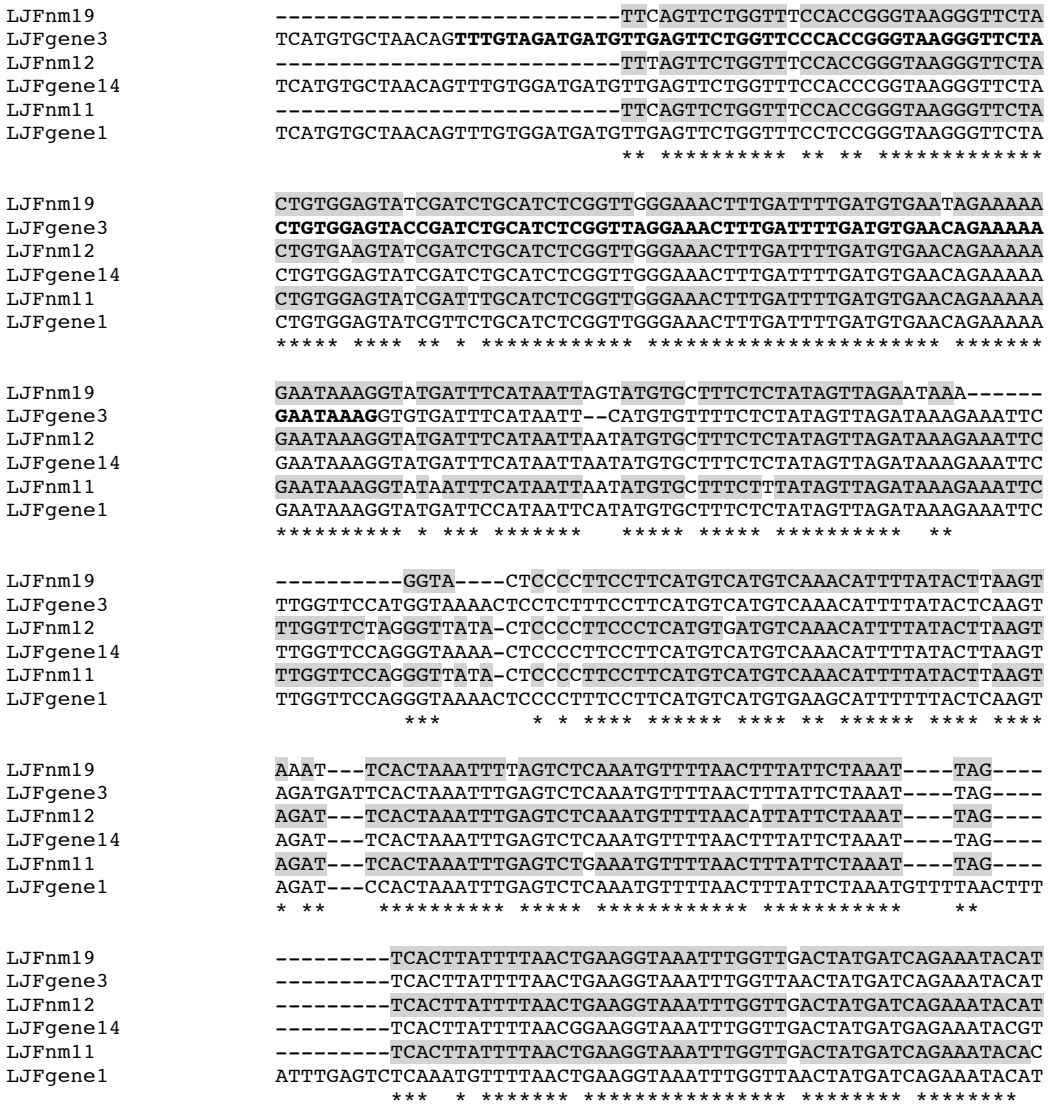


Figure 3.30. Partial multiple alignment output of sequences from chromosomes 19, 11, and 12 containing LJFnm members against LJFgene3, LJFgene14, and LJFgene1 beginning in intron 5 of LJFgene family members and extending to 3' most nucleotides of non-coding chromosomal sequences that display strong identity with sequences of the LJFgene family members. Bold type represents coding sequence of LJFgene3; grey highlights indicate identity of LJFnm genomic sequence with LJFgene3 genomic sequence.

```
LJFnm19             TGATATTTTTTAATTGGTAGAGATAAAGAATATTTTTTATGTACAATAAAGAGAGTATTT
LJFgene3            TGACATTTTTTAATTGGTAGAGATAAAGAATATTTTTTATGTACAATAAAGAGAGTATTT
LJFnm12             TGATATTTTTTAATTGGTAGAGATAAAGAATATTTTTTGATGTACAATAAAGAGAGTACTT
LJFgene14           TGATATTTTTTAATTGGTAGAGATAAAGAATATTTTTTTATGTACAATAAAGAGAGTATTT
LJFnm11             TGATATTTTTTAATTGGTAGAGATAAAGAATATTTTTTTATGTACAATAAAGAGAGTATTT
LJFgene1            TAACA-------------AGAGTTGAAGAATATTTTTTTATGTACAATAAAGAGAGTATTT
                    * * *             **** * ************ ****************** **

 LJFnm19             ACTCCAGAGGATGCAAATCCCTTACTAAATATTTTTGTGATGAAAAATCTTGGTTGCTGA
LJFgene3            ACTCCAGAGGATGTAAATCCCTTGCTAAATATTTTTGTGATGAAAAATCTTGGTTGTTGA
LJFnm12             ACTCCAGAGGATGCAAATCCCTTACTAAATATTTTTGTGATGAAAAATCTTGGTTGCTGA
LJFgene14           ACTCCAGAGGATGCAAATCCCTTACTAAATATTTTTGTGATGAAAAATCTTGGTTGCTGA
LJFnm11             ACTCCAGAGGATGCAAATCCTTTACTAAATATTTTTGTGATTAAAAATCTTGGTTGCTGA
LJFgene1            GCTCGAGAGAATGTAAATCCTTTTCTAAATATTTTTGTGATGAAAAATAATGGTTGCTGG
                     *** **** *** ****** ** ***************** ******  ****** **

LJFnm19             CAGGCATTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTCAAGATACCATATGATTA
LJFgene3            CAG**GCACTGCGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCAAGACACCATATGA**TGA
LJFnm12             AAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTCAAGATACCATATGATTA
LJFgene14           CAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTCAAGATACCATATGATTA
LJFnm11             CAGGCACTGAGACAAGAGTTAGAGAAGAAAGGATGGACATCTCAAGATACCATATGATTA
LJFgene1            CAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCAAGACACCATATGATGA
                     ***** ** ********** ************** ********** ********** *

LJFnm19             ATAAACTCAGGCAGAATTAACATCAACATCTAAGCAAATATTATTTCATATACTTTGTGA
LJFgene3            ATAAACTCAGGCAGAATTAACATCAGCATCTAAGCAAATATTATTTCATATACTTTGTGA
LJFnm12             ATAAACTCAGGCAGAATTAACATCAGCATCTAAGCAAATATTATTTCATATACTTTGTGA
LJFgene14           ATAAACTCAGGCTGAATTAGCATCAGCATCTAAGCAAATATTATTTCATATACTTTG-GA
LJFnm11             ATAAACTCAGGCAGAATTAACATCAGCATCTAAGCAAATATTATTTCATATACTTTGTGA
LJFgene1            AAAAACTTAGGCAGAATTCACATCAGCATCTAAGAAAATATTGTTTCATATACATTGTAA
                    * ***** **** *****  ***** ******** ******* ********** ***  *

LJFnm19             CCTTGTATACTTTTGTATTAGATACAAA-TCACACAGGATCATTTCAAGCAAACTTTTCT
LJFgene3            CCTTGTATACATTTGTATTAGATACAAA-TCTCACAGGATCATTGAAAGCAAACTTTTCT
LJFnm12             CCTTGTATACTTTTGTATTAGATACAAA-T-------------TGCAAGCAAACTTTTCT
LJFgene14           CCTTGTATACTTTTGTATTAGATACAAA-TCGCACAGGATCATTGCAAGCAAACTTTTCT
LJFnm11             CCTTGTATACTTTTGTATTAGATACAAA-TCGCACAGGATCATTGCAAGCAAACTTTTCT
LJFgene1            CCTTGTATACTTTTGTATTAGATACAAAATCTCACAAGATCATTGAAAGCAAACTCTTCA
                    ********** ***************** *             *  ********* ***

LJFnm19             TAGATTTTAGGAATTGTAGAGAAATCATTGAGA-CAATACTTTAAACTCTC--GGGGAAG
LJFgene3            TTGATTATTGGAATTGTAGAGAAATCATTGAGAACAGTACTTCAAACTCTC--GGGGAAG
LJFnm12             TAGATTTTTG-------------------------------------------------
LJFgene14           TAGATTTTTGGAATTGTAGAGAAATCATTGAGAACAGTACTTCAAACTCTCTCGGGGAAG
LJFnm11             TAGATTTTTGGAATTGTAGAGAAATCATTGAGAACAATACTTCAAACTCTC--GGGGAAG
LJFgene1            T-GATTATTGGAATTGTAGA--AATGATTGAGAACAGTACTTCAAACTCTCG-GGGGAAG
                    * **** * *

LJFnm19             GAATGGAATGAAGACCTTG-----------------------------------------
LJFgene3            GAATGAAATGAAGACCTTGCCCCATATCCTTCTTCAAGTTCATTAATTGGTCCGCTTATT
LJFnm12             -----------------------------------------------------------
LJFgene14           GAATGAAATGAAGACCTTGCGCCATATC-TTCTTCAAGTTCATTAATTGGTCCACTTATT
LJFnm11             GAATGAAATGAAGACCTTG-----------------------------------------
LJFgene1            GAATGAAATGAAGATGTTACC---------------------------------------
```

Figure 3.30. Partial multiple alignment output of sequences from chromosomes 19, 11, and 12 containing LJFnm members against LJFgene3, LJFgene14, and LJFgene1 beginning in intron 5 of LJFgene family members and extending to 3' most nucleotides of non-coding chromosomal sequences that display strong identity with sequences of the LJFgene family members. (CONT.)

Figure 3.31 displays a dot plot matrix generated using the genomic sequence of LJFgene3 and a 4000 nucleotide segment of chromosome 19 that contains LJFnm19 as

input. The position of LJFnm19 relative to the LJFgene3 model is confirmed to be near the 3' end of LJFgene3.

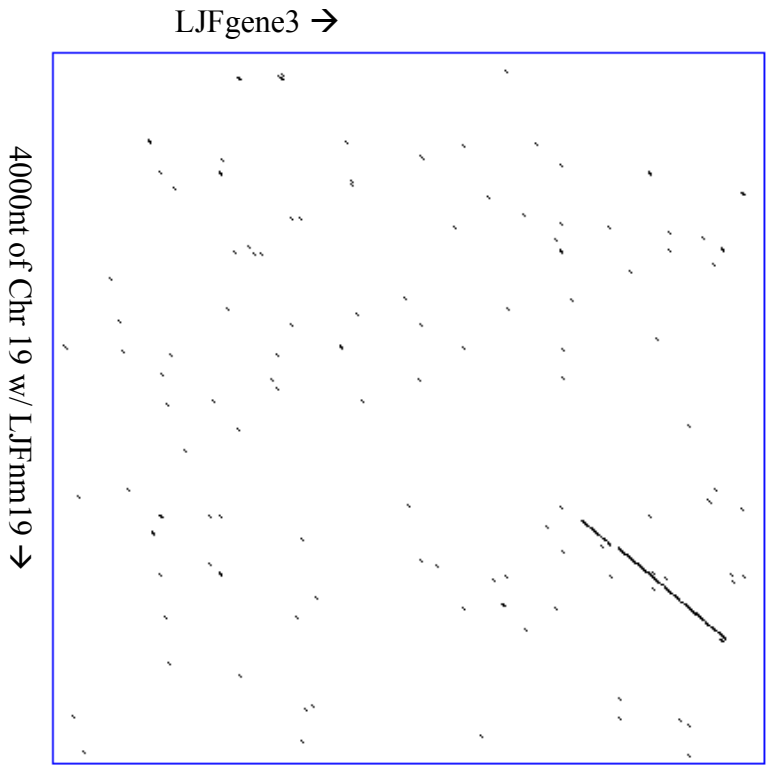


Figure 3.31. Dot plot matrix of LJFgene3 genomic sequence (x-axis) vs. 4000nt segment of chromosome 19 containing LJFnm19 (y-axis).

3.5.4. MicroRNA Prediction. Of the sequences submitted to the prediction tool, LJFnm12 and LJFnm19 were predicted to contain miRNA precursors. Figure 3.32 illustrates the position and specific sequence predicted to be a miRNA precursor within the conceptually transcribed mRNA sequence of LJFnm19 and LJFnm12. LJFnm11 was not predicted to contain a miRNA precursor. Additionally, the two predicted miRNA precursors do not correspond to the same section of sequence on respective LJFnm sequences.

(A)
LJFnm19:
UUCAGUUCUGGUUUCCACCGGGUAAGGGUUCUACUGUGGAGUAUCGAUCUGCAUCUCG
GUUGGGAAACUUUGAUUUUGAUGUGAAUAGAAAAAGAAUAAAGGUAUGAUUUCAUAAU
UAGUAUGUGCUUUCUCUAUAGUUAGAAUAAAGGUACUCCCCUUCCUUCAUGUCAUGUC
AAACAUUUUAUACUUAAGUAAAUUCACUAAAUUUUAGUCUCAAAUGUUUUAACUUUAU
UCUAAAUUAGUCACUUAUUUUAACUGAAGGUAAAUUUGGUUGACUAUGAUCAGAAAUA
CAUUGAUAUUUUUUAAUUGGUAGAGAUAAAGAAUAUUUUUUAUGUACAAUAAAGAGAG
UAUUUACUCCAGAGGAUGCAAAUCCCUUACUAAAUAUUUUUGUGAUGAAAAAUCUUGG
UUGCUGACAGGCAUUGAGACAAGAGUUGGAGAAGAAAGGAUGGACAUCUCAAGAUACC
AUAUGAUUAAUAAACUCAGGCAGAAUUAACAUCAACAUCUAAGCAAAUAUUAUUUCAU
AUACUUUGUGACCUUGUAUACUUUUGUAUUAGAUACAAAUCACACAGGAUCAUUUCAA
GCAAACUUUUCUUAGAUUUUAGGAAUUGUAGAGAAAUCAUUGAGACAAUACUUUAAAC
UCUCGGGGAAGGAAUGGAAUGAAGACCUUG

(B)
LJFnm12:
UUUAGUUCUGGUUUCCACCGGGUAAGGGUUCUACUGUGAAGUAUCGAUCUGCAUCUCG
GUUGGGAAACUUUGAUUUUGAUGUGAACAGAAAAAGAAUAAAGGUAUGAUUUCAUAAU
UAAUAUGUGCUUUCUCUAUAGUUAGAUAAAGAAAUUCUUGGUUCUAGGGUUAUACUCC
CCUUCCCUCAUGUGAUGUCAAACAUUUUAUACUUAAGUAGAUUCACUAAAUUUGAGUC
UCAAAUGUUUUAACAUUAUUCUAAAUUAGUCACUUAUUUUAACUGAAGGUAAAUUUGG
UUGACUAUGAUCAGAAAUACAUUGAUAUUUUUUAAUUGGUAGAGAUAAAGAAUAUUUU
UGAUGUACAAUAAAGAGAGUACUUACUCCAGAGGAUGCAAAUCCCUUACUAAAUAUUU
UUGUGAUGAAAAAUCUUGGUUGCUGAAAGGCACUGAGACAAGAGUUGGAGAAGAAAGG
AUGGACAUCUCAAGAUACCAUAUGAUUAAUAAACUCAGGCAGAAUUAACAUCAGCAUC
UAAGCAAAUAUUAUUUCAUAUACUUUGUGACCUUGUAUACUUUUGUAUUAGAUACAAA
UUGCAAGCAAACUUUUCUUAGAUUUUUG

Figure 3.32. Results of microRNA prediction by web-based tool using fixed-order hidden markov model. (A) LJFnm19 transcribed genomic sequence. (B) LJFnm12 transcribed genomic sequence. Red capital letters indicate predicted mature miRNA sequence.

3.5.5. Promoter Element Identification. Sequences upstream of the 5' end of the LJFnm sequences were submitted to PLACE for identification of promotor elements associated with transcription of a gene. Figures 3.33, 3.34, and 3.35 display the findings of this promoter element search for sequences LJFnm19, LJFnm12 and LJFnm11, respectively.

caattatgctgcataaaaattatgatacagttgaagcaagtagaaataat
tactatagttttatcaaaatgaattatgcagttatcaattatgtaaaatt
aatttcgctgaaaaattaattttattaccgtgaatccaaacacatactaa
aataattacttggcaatggctcctctataagagaggagtgtaaagcagga
atttttactcatgtgcacaaatccgttggggtgttttgggatatgattt
tttggatttaaggacttgcatatgggtggtgttttattggattcccattc
acaagtcattttgcatacaaactatgttgtacaacagatgtgtataccct
atatttaacattagtttgacagcagagaccttctttaattacatcagcag
caacccaaaaataaacccccacacacctgccctttcagaaaggtaaacac
taataatacaaactctggaactaatcacaaggcataaaataattcttgca
actacacataaataaagtaacttaaatatgcagaaaacatttgtgcagga
acttcaacacctgaggagtgagcgatgaagactacacaaaattggtggac
atatactcatttgggatgtgtgtgttagagctggtgatagtatagattcc
ttatagtgaatgtgacaatgttgataagatatacaagaaggtgtcttctg
gagttagacctgctgccttgaacaaggtcaaagatcctgaggttaaggct
ttcattgagaagtgccttgctcagccaagggctaggccttctgcagttga
gcttctcaaagatcctttctttgatgagattgatgatgatgatgacaaaa
atgatgattgttcttgttcatatcaataagaataatatttcttacagtgtt
ggtttttaatttgattcacctacctttcagttctggtttccaccgggt
aagggttctactgtggagtatcgatctgcatctcggttgggaaactttga
ttttgatgtgaatagaaaaagaataaaggtatgatttcataattagtatg

Figure 3.33. Cis-acting elements (highlighted blue) located within a 1Kbp segment of sequence from chromosome 19 that includes LJFnm19 (highlighted gray).

atgtatttttggctattaattaaacaaaatcaaataagaaaagaaaatt
acctgaatatcctcccaaatcaggtctttctgagcaccagggacctactt
ctagttctcgtaggtgacgtccaccttatcacacgccacaatccccaaat
atgttcttaatttcttcctgtggggaccgtcggccttgccagtagcagga
tcaacatggaccactggtctctctgcaccaggtggtctagtagacaacga
tcgtagacatgaggctttgcgtgtccgcttcacggtagacggcgaagccg
atgcgtcggaaggaggaggaggaggaggaggggggaggggcaggtggaga
agccatgatcctttataaacaataacaacaacatacaaaatgtaaatt
tcgattacagacaaatatcaacatctagagttctttatgtaagtaatctg
gaatagtggagttcttacagtgttggtttttaatttgattcacctacct
ttttagttctggtttccaccgggtaagggttctactgtgaagtatcgatc
tgcatctcggttgggaaactttgattttgatgtgaacagaaaaagaataa
aggtatgatttcataattaatatgtgctttctctatagttagataaagaa
attcttggttctagggttatactccccttccctcatgtgatgtcaaacat
tttatacttaagtagattcactaaatttgagtctcaaatgttttaacatt
attctaaattagtcacttattttaactgaaggtaaatttggttgactatg
atcagaaatacattgatatttttttaattggtagagataaagaatattttt
gatgtacaataaagagagtacttactccagaggatgcaaatcccttacta
aatatttttgtgatgaaaaatcttggttgctgaaaggcactgagacaaga
gttggagaagaaaggatggacatctcaagataccatatgattaataaact
caggcagaattaacatcagcatctaagcaaatattatttcatatactttg

Figure 3.34. Cis-acting elements (highlighted blue) located within a 1Kbp segment of sequence from chromosome 12 that includes LJFnm12 (highlighted gray).

aggaatacctagcttgacatataaccttagaatgcaatttcaggccaat
ttggagaagtatatcaagcaacttctctaaattagggtttggcaacaatt
atgggctagaaccattgtgcatacatgattcttggcacaccataatttat
ggttctagacttatacgatgaagactacacaaaattggtggacatatact
catttgggatgtgtgtgttagagctggtgacagtagagattccttatagt
gaatgtgacaatgttgataagatatacaaaaaggtgtcttctggagttag
acctactgccttgaacaaggtcaaagatcctaaggttaaggctttcattg
agaagtgccttgctcagccaagggctaggccttctgcagctgagcttctc
agagatcctttctttgatgagattgttgatgatggtgacgaaaatgatga
ctgttcttgttcatatcaataagaataatatttcttacagtgttggttttt
taatttgattcacctacctttcagttctggtttccaccgggtaagggtt
ctactgtggagtatcgatttgcatctcggttgggaaactttgattttgat
gtgaacagaaaaagaataaaggtatgatttcataattaatatgtgctttc
tctatagttagataaagaaattgttggttccagggttatactccccttcc
ttcatgtcatgtcaaacattttatacttaagtagattcactaaatttgag
tctcaaatgttttaactttattctaaattagtcacttattttaactgaag
gtaaatttggttgactatgattagtaatacattgatatttttaattggt
agagataaagaatatttttgatgtacaataaagagagtatttactccaga
ggatgcaaatcccttactaaatatttttgtgatgaaaaatcttggttgct
gacaggcactgagacaagagttggagaagaaaggatggacatctcaagat

Figure 3.35. Cis-acting elements (highlighted blue) located within a 1Kbp segment of sequence from chromosome 11 that includes LJFnm11 (highlighted gray).

4. DISCUSSION

4.1. CHOICE OF GENE FAMILY AND IDENTIFICATION OF MEMBERS

The selection of a single family of interest for the focus of this study was conducted by manually searching through multiple databases to identify a family that could meet predetermined criteria. The criteria by which the family was chosen are:

- 10 or fewer gene members
- shows expansion in *Glycine max* relative to other plant species
- unknown functional annotation
- each potential family member must point to all other potential family members during a BLAST search of the genome.

The purpose of restricting the number of genes and considering only families composed of immediate members was to increase the likelihood of correctly identifying and incorporating the complete family. Choosing a family that has an expanded number of genes in soybean relative to other plant species but no known functional data was a matter of personal interest. The hope was that if the family is expanded in soybean, it is an indication of increased or novel function and that through in depth analysis of the coding sequence a likely gene product form and function could be predicted. A record of the partial results of the search by criteria can be located in Table 3.1 in Section 3.1.1.

4.2. GENE STRUCTURE PREDICTION AND EST EXPRESSION ANALYSIS

Once focus was narrowed to a single gene family of interest, a chromosome map (Figure 3.2) was generated for each gene within the family. To determine DNA composition at the location of gene family members on their respective chromosomes, the chromosome maps were then compared to a genomic landscape analysis of the 20 soybean chromosomes reported in Schmutz et al. [9] The analysis reported the percent composition of major DNA elements including transposons, retrotransposons, centromeric DNA, and coding DNA at each position of the *Glycine max* chromosomes. Composition was calculated using 0.5 Mb (500 Kbp) windows with a 0.1 Mb (100 Kbp) shift. The composition of the 100 Kbp regions of the chromosome that gene family members lie within are fairly similar. All gene family members lie near the ends of the chromosomes where coding sequence composition is roughly between 20 and 40 percent, transpsosons and retrotransposons make up a small fraction of the sequence, and the majority of the sequence is unclassified DNA.

FgenesH, GenomeScan, and Augustus were employed to predict gene models for family members [9, 48]. These predicted models were then uploaded into the DNA Subway annotation tool for refinement. The final model determination took into account predicted models, known consensus data concerning intron/exon border sequences, open reading frame calculations, and EST data. The best evidence for the existence of genes is empirical, and that can be found in the form of expressed sequence tags.

Most of the ESTs available for this gene family (Appendix F) agree with the arrangement of introns and exons predicted by the algorithm-generated models. It was

necessary, however, to align the EST nucleic acid sequence from NCBI to the genomic regions using the annotation editor. Since ESTs are partial cDNA's, they reflect a sequence from which introns have been spliced out. Alignments made by the annotation editor do not break across introns at correct splice junctions. Consensus data was used to verify intron/exon borders.

The gene family member on chromosome 3 has the highest quantity of ESTs of all family members at seven ESTs. Genes on chromosomes 14 and 9 have three ESTs each and genes on chromosomes 1 and 8 have no representative ESTs (Figure 3.2). Due to the majority of the ESTs (evidence of expression) and completeness of the model, LJFgene3 is considered most likely to produce a functional gene product and therefore the model by which all other gene family members are compared to for evolutionary analysis.

Genes on chromosomes 3 and 14 share >85% sequence identity at the nucleotide level. The EST data available for the gene family member on chromosome 14 disagreed with the algorithm-predicted models on the intron/exon arrangement at the 5' end. All three gene-predicting computer programs with independent algorithms produced models for LJFgene14 that have features most closely resembling LJFgene3. While algorithms predict a model with 7 exons and 6 introns, EST evidence indicates a model with 6 exons and 5 introns. Other evidence, such as PASA-assembled EST data, RNA-seq data, and transcription-level expression data, supports the LJFgene14 model shown in Figure 4.1.

The only difference between the two models is the inclusion of the segment of DNA that corresponds to intron 2 in the algorithm-predicted model into exon 2 of the

EST-derived model. In other words, the sequence spanning exon 2, intron 2, and exon 3 of the predicted model is exon 2 of the EST-based model (Figure 4.1).

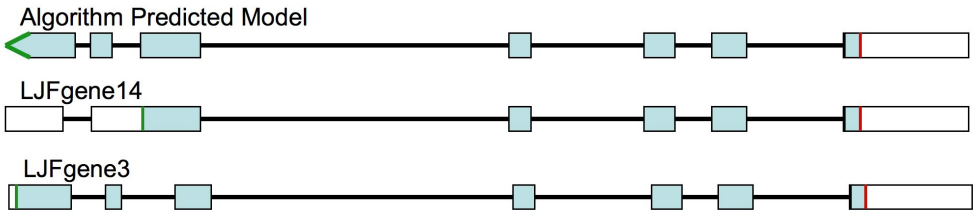


Figure 4.1. Algorithm-predicted model of gene family member on chromosome 14 vs. LJFgene14 (model generated using EST evidence). LJFgene3 is included as a reference.

4.3. EVOLUTIONARY ANALYSIS

In either model it can be seen that the position of the start site is affected. In the algorithm-generated model, a start methionine cannot be defined at the 5' end using the reading frame that corresponds to the most likely coding arrangement. In the EST-generated model, the start site is located in the middle of exon 2. Using LJFgene3 as the basis for comparison, a pairwise alignment of nucleic acid sequence (Figure 4.2) in the region in question can illuminate the causal dissimilarities between models.

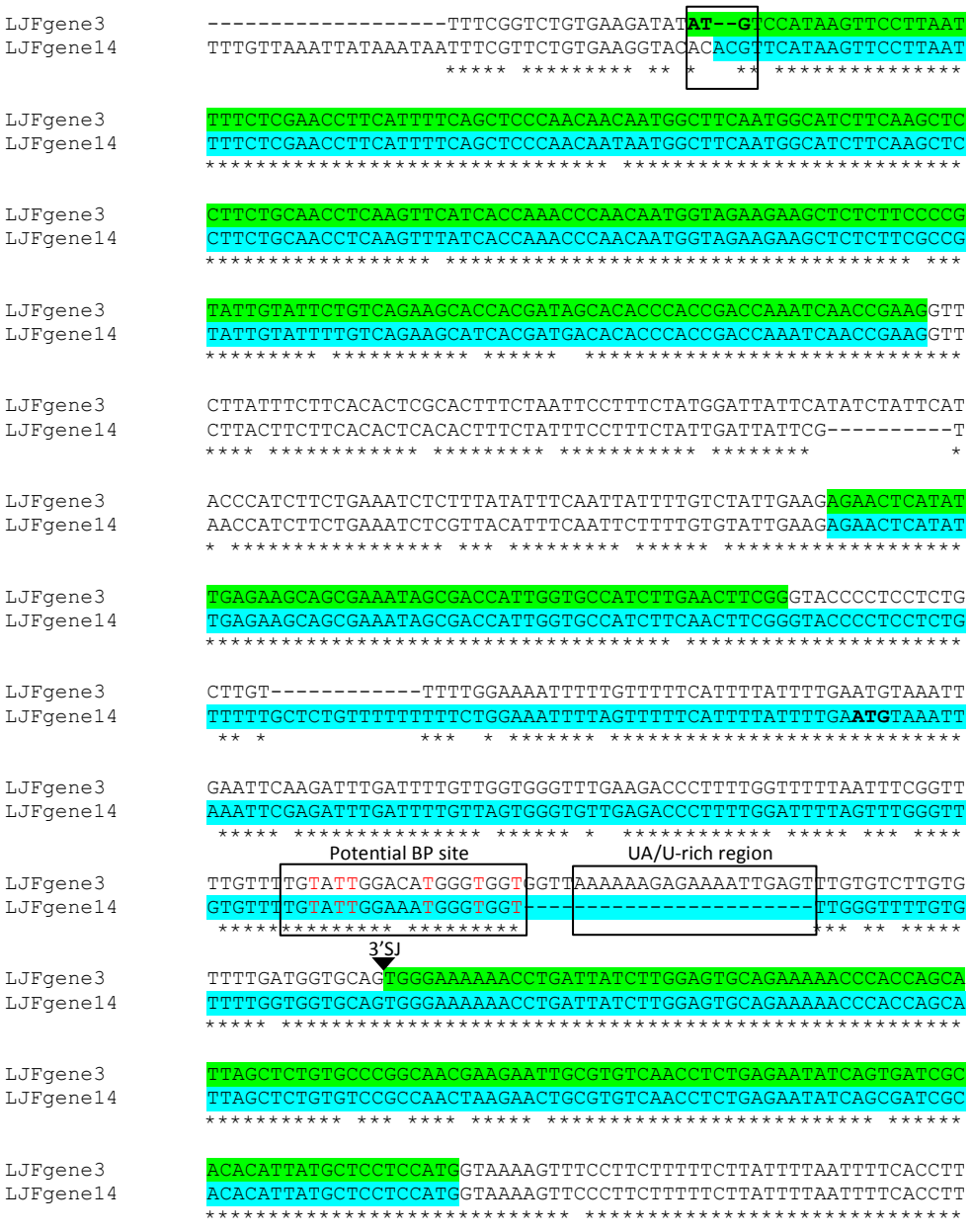


Figure 4.2. Pairwise alignment of 5' end of LJFgene3 and LJFgene14 nucleic acid sequences. (Nucleic acids corresponding to LJFgene3 model are highlighted in green. Nucleic acids corresponding to LJFgene14 model are highlighted in blue. Start ATG sequences for each model are in bold face type. Possible key alternative splicing sites are indicated by boxes and a potential branch point adenosine is indicated by red text.)

The nucleotide sequence of LJFgene14 that is directly aligned with the LJFgene3 start ATG differs by 3 nucleotides. A transition point mutation has occurred in LJFgene14 at the T position of LJFgene3 and a two base pair insertion-deletion (indel) has disrupted the second and third position of the ATG sequence. It cannot be known which mutation occurred first or in which direction, only that they exist in this way in their current forms. This explains why neither gene model for the gene family member on chromosome 14 has a start site matching LJFgene3.

Evidence for intron 2 retention and the location of the start site in exon 2, as seen in the EST-derived model, can also be found by examination of the pairwise alignment of LJFgene3 and LJFgene14 found in Figure 4.2. For intron 2 to be spliced out of the transcript, evidence should exist in the form of sequences corresponding to known polypyrimidine (poly-p) tracts, branch point consensus sequences, and/or UA/U-rich regions. The function of these elements depends on presence and location. If a branch point is present, the UA-rich regions may act as a poly-p tract. In the absence of a branch point, the UA-rich regions may act as intronic splice elements [28]. In either case, the UA-rich regions must have a minimum U content and be separated by a minimum distance to maintain proper splice efficiency. Some plant genes such as the potato invertase (invGF) gene contain poly-p tracts in their introns that consist of long strings of consecutive U's (11 in the case of the invGF gene), other dicot genes cannot be supported by a single group of consecutive U's; rather, they require multiple, smaller groups of U's. A mutational study of the invGF gene introns provides evidence that two groups of four U's each that are spaced 3 C's apart is the optimal arrangement [31].

Since the poly-p tract of the transcript is usually a segment of tandemly repeating U's, the DNA sequence that corresponds to it should contain tandem repeats of A's. Since a UA/U-rich region would also resemble this, it becomes difficult to distinguish one from the other. Regardless, one should exist within the range of 17 to 40 nucleotides upstream of the 3' splice junction (SJ) [25]. This region of repeating A's exists in intron 2 of LJFgene3 from -45 to -33 nucleotides (AAAAAAGAGAAAA) upstream of the 3' SJ. This sequence meets the requirements of U-rich elements acting as a poly-p tract—two groups of four or more U's separated by three pyrimidines (CUC in this case). The section of sequence of LJFgene14 that directly aligns with LJFgene3 at this position is represented by gaps, indicating that an indel has occurred in this position. The mutation is most likely a deletion based on the phylogenetic relationship. In an alignment of LJFgene3 and LJFgene1, this segment of sequence is identical. A single deletion occurrence in an ancestor is more parsimonious than a deletion followed by reinsertion of sequence. Logic dictates that without a site for the binding of splicesome formation-mediating proteins, the splicesome would be unable to form and the intron would not be able to be spliced out. This presents a reasonable explanation for the retention of what might have been intron 2. As it turns out, the intron retention provides the first possible ATG in the open reading frame. In addition, it has been demonstrated that dicot plants require a minimum intronic AU content of 59% in order to maintain efficient splicing [29]. The region of sequence in LJFgene14 that corresponds to intron 2 of gene LJFgene3 has 58.65% AU content (intron 2 of LJFgene3 has 65.5% AU content). Although this

number is borderline, it could have an impact on splice efficiency and be a possible alternate explanation for the intron retention.

Branch point sequences exactly matching the experimentally verified branch point consensus sequences were not found in the introns of gene family members LJFgene3, LJFgene14, or LJFgene1. This does not mean that the branch points do not exist, merely that they cannot be identified using this evidence. Even so, in the absence of a branch point, the splice machinery still requires a UA/U-rich region for recruitment and assembly.

The establishment of the correct gene model for the gene family member on chromosome 14 is important to understanding the evolutionary history of this family and how it affects the assembly of a phylogenetic tree. Two trees were generated for this family, one using the original gene models (which were based on conceptually translated peptides representing the algorithm generated models and therefore characteristic of the nucleotide sequences) and the final gene models (which were based on the conceptually translated peptide sequences that correspond to evidence of expression). Both trees place LJFgene9 and LJFgene8 as diverging separately from the clade including LJFgene1, LJFgene14, and LJFgene3. The trees disagree on the relationship between LJFgene1, LJFgene14, and LJFgene3. Figure 4.3 illustrates this difference.

Gene Comparisons (by chromosome)	Synonymous Substitution Rate of ORIGINAL gene models	Synonymous Substitution Rate of FINAL gene models
3 vs 14	.1103	.1781
3 vs 1	.1203	.1303
3 vs 8	.2775	.2991
3 vs 9	.6273	.5512

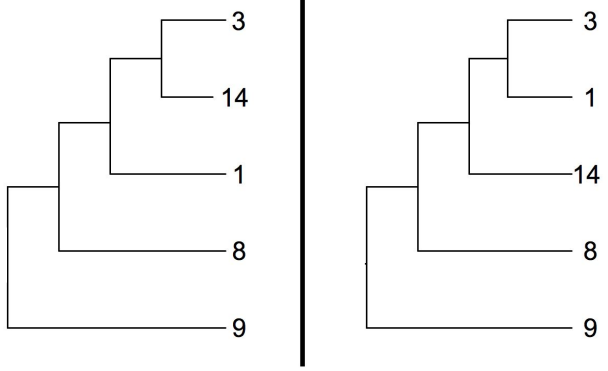


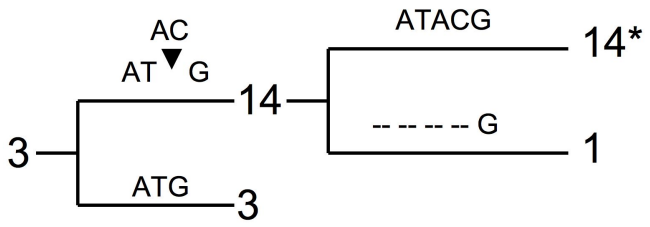
Figure 4.3. Comparison of the synonymous substitution rate and resulting phylogentic differences between original gene models and final gene models.

The phylogenies differ because the synonymous substitution rate is calculated by a program that uses the codon alignment that is based on conceptually translated peptide sequences aligned with nucleotide sequences. In the amino acid sequence of LJFgene14, it reflects the placement of the start site due to intron retention in the final gene model. The final models, therefore, are a better representation of the gene product and the original model remains the best representation of the evolutionary relationship at the DNA level.

In the final phylogenetic model, the divergence pattern of LJFgene3, LJFgene14, and LJFgene1 is rearranged. The synonymous substitution rate between

LJFgene3 and LJFgene14, a smaller number (indicating a briefer time lapse since divergence) in the original calculations, becomes a larger number than the synonymous substitution rate for LJFgene1 for the final models. Though not listed in Figure 4.3 above, a comparison of gene family members on chromosomes 14 and 1 yields a synonymous substitution rate of 0.1417 (Table 3.5). LJFgene1 has nearly the same synonymous substitution rate when compared to LJFgene14 (0.1417) as it does when compared to LJFgene3 (0.1303); however, the rate is still lower between LJFgene1 and LJFgene3 indicating fewer mutations have accumulated and therefore inferring a shorter amount of time since divergence.

Depending on the evidence used, two possible evolutionary trajectories can be considered when dissecting the relationship between LJFgene3, LJFgene14, and LJFgene1. According to the synonymous substitution rate as calculated for the original gene models (it has already been established that these calculations represent the evolution of the genes at the nucleotide level), LJFgene14 is more closely related to LJFgene3 than LJFgene1. According to this reasoning, the phylogeny should appear as it does in Figure 4.4.



* The final genomic sequence of this location is actually ACACG. A Point mutation occurs at the T position; however, it is not reflected in this model since the timing of the occurrence cannot be determined.

```
LJFgene3        GAAGATATAT--GTCCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAAC
LJFgene14       GAAGGTACACACGTTCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAAC
LJFgene1        GAAGATAC----GTTCATAAGTTCCTTAATTTTCTTGAACCTTTATTTTCAGCTCCCAAC
```

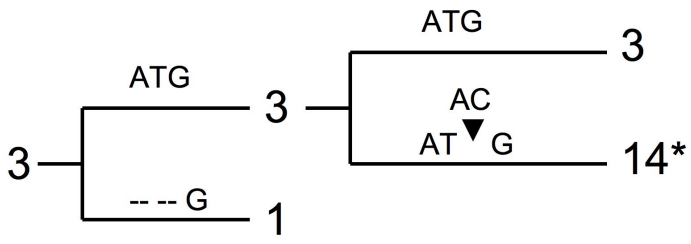
Figure 4.4. Phylogenetic relationship and mutations occurring in functional start site between LJFgene3, LJFgene14, and LJFgene1: Scenario 1.

If LJFgene14 diverged before LJFgene1 from the common ancestor of LJFgene3, LJFgene14, and LJFgene1, the disruption of the functional start site as seen in Figure 4.4 would have begun with a 2 base pair (AC) insertion between the T and G in the common ancestor to LJFgene14 and LJFgene1. A 4 base pair deletion of the ATAC preceding the G would have occurred in the lineage leading to LJFgene1.

The absence of expression data forces the algorithm-predicted model of LJFgene1 to be compared to two models for which expression evidence exists. If the model for LJFgene1 is not an accurate reflection of the physical gene product (if one is produced at all), the synonymous substitution rate could be inaccurate and a phylogeny with synonymous substitution rates resembling those of the final gene models might be the true lineage.

If LJFgene1 diverged before LJFgene14 from the common ancestor of LJFgene3, LJFgene14, and LJFgene1, the disruption of the functional start site as seen in Figure 4.5 would have begun with a 2 base pair deletion of the AT in LJFgene1. A 2

base pair (AC) insertion would have occurred between the T and G in the lineage of LJFgene14 after its divergence from the common ancestor of LJFgene3 and LJFgene14.



* The final genomic sequence of this location is actually ACACG. A Point mutation occurs at the T position; however, it is not reflected in this model since the timing of the occurrence cannot be determined.

```
LJFgene3        GAAGATATAT--GTCCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAAC
LJFgene14       GAAGGTACACACGTTCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAAC
LJFgene1        GAAGATAC----GTTCATAAGTTCCTTAATTTTCTTGAACCTTTATTTTCAGCTCCCAAC
```

Figure 4.5. Phylogenetic relationship and mutations occurring in functional start site between LJFgene3, LJFgene14, and LJFgene1: Scenario 2.

Based on the understanding of the clock-like patterns that synonymous substitutions produce and the time frames in which researchers believe whole genome duplications occurred in the history of *Glycine max*, the divergence patterns that have led to the modern composition of genes in this family can be tentatively associated with major genome-impacting events in the *Glycine max* history. The synonymous substitution rates (Table 3.5) for LJFgene3, LJFgene14, and LJFgene1 all fall into a commonly accepted range corresponding to the 13 Mya soybean whole genome duplication (WGD) [9]. It is possible that these closely related genes could have resulted from the WGD being followed by a segmental duplication that produced 3

lineages. The synonymous substitution rates for LJFgene8 also fall within the accepted range for the 13 Mya WGD; however, the rate is different enough to consider the possibility that this gene might have resulted from a segmental duplication prior to the WGD. The synonymous substitution rate for LJFgene9 falls within the acceptable range of scores corresponding to a 59 Mya WGD that gave rise to the legume clade. Although these values can be matched to major duplication events, one must be careful not to over interpret the meaning of the values. The synonymous substitution rates, as outlined by Schmutz et al. [9], are generated using the entire soybean genome in comparison to itself. The synonymous substitution rates generated for the LJFgene family are a result of comparing the genes within the family to each other, not to the entire genome. The Ks/ps values generated by this research can only be interpreted as evidence for the chronological difference between the gene sequences in this family as measured by the rate of synonymous substitution accumulation. Far more synonymnous mutations have accumulated between the sequences of LJFgene3 and LJFgene9 than any other family member comparison indicating that they diverged in more distant past. Conversely, the sequences of LJFgene3 and LJFgene1 have the lowest rate of synonymous mutations accumulation and are therefore interpreted as having diverged from each other most recently.

Given that the synonymous substitution rate calculations include a certain amount of error, and that the values are so close between LJFgene3, LJFgene14, and LJFgene1, this evidence could not be used to conclusively define the divergence pattern of these three genes. An analysis of homology between the *Glycine max* chromosomes conducted by Schmutz et al. [9] based on the presence of specific

centromeric repeat regions indicates that chromosomes 1 and 3 are homologous and more likely to have originated from the 13 Mya soybean WGD. Combined with the neighbor gene analysis (intended to analyze syntenic regions on the chromosomes flanking gene family members) results that showed a possible syntenic pattern between chromosomes 1 and 3, this suggests that LJFgene1 might have originated from the soybean-specific WGD. No evidence exists in the data collected to discern whether LJFgene14 originated from this event. The Schmutz et al. [9] homology analysis previously mentioned indicates that chromosome 14 shares more homology with chromosomes 8 and 9 based on the presence of centromeric repeats.

With the exception of the 5' boundary of the first exon, LJFgene1 exon boundaries match LJFgene3 exon boundaries exactly. The sequences within these exons are very nearly the same, meaning the gene product of LJFgene1 is nearly the same as the gene product of LJFgene3, aside from LJFgene1 theoretically producing a truncated protein. This is supported by the multiple alignment of conceptually translated peptides located in the Results Section as Figure 3.7. The synonymous substitution rate for the comparison of LJFgene3 and LJFgene1 is low because of this high degree of sequence similarity within the coding regions of these genes, yet the genomic sequence tells a different story. Two sizeable regions of sequence exist in LJFgene1 that do not exist in LJFgene3. These indels occur in the middle of introns 3 and 4. They do not exist in LJFgene14 either, which strongly indicates these extra sequences have been inserted into LJFgene1. The insertion in intron 3 is only 184 nucleotides long and falls nearer to the 5' splice junction. The insertion in intron 4 is more sizeable, nearly a thousand extra nucleotides, and falls closer to the 3' splice

junction. Figure 4.6 illustrates the shift in exon proximity due to increased intron length. It appears to insert enough nucleotides upstream from this boundary not to disrupt the presence of vital splice elements, however, an insertion this large could potentially make the gap between exons large enough to have an effect on splice efficiency. Without the necessary expression evidence, it cannot be determined whether this anomaly would produce a transcript with extra sequence or not. Perhaps it is a clue as to why no ESTs have been linked to this gene model.

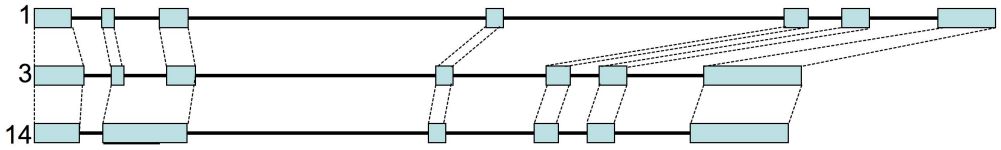


Figure 4.6. Gene models of LJFgene3, LJFgene14, and LJFgene1 displaying close approximation of exon size and shift in exon proximity in LJFgene1 due to intron length.

Given that the largest degree of expression was linked to LJFgene3, The assumption has been made that this gene is the most likely candidate to produce a functional gene product. Differences in sequence, structure, and expression lead to the conclusion that the remaining four gene family members are pseudogenes, two of which (LJFgene14 and LJFgene9) are still producing mRNA but not a functional protein. It is interesting that LJFgene9, which has the lowest percentage of sequence similarity to LJFgene3, still produces enough mRNA to account for 3 ESTs; yet LJFgene8, which resembles LJFgene3 to a much higher degree, has no ESTs in the library. Due to high level of sequence variation between the 5’ and 3’ ends of the gene model for LJFgene9 and all other gene family members, the assignment of the gene to

this family was brought into question. The nucleic acid sequences of the segments of LJFgene9 that are non-congruent with the rest of the family were submitted as a BLAST search against the *Glycine max* genome in an attempt to determine whether the gene model resulted from a rearrangement or belongs to another gene family. No matches were recorded to any gene models outside of this family.

In order to better understand the full coding capacity of these genes and their flanking sequences, a multiple sequence alignment was performed that extended the query sequence beyond each model's boundaries far enough to compare a stretch of DNA sequence for each model that is at least as long as LJFgene3 (Figure 3.7). The conceptually translated amino acid sequences of the extended genes were aligned and compared. A summary of sequence extension requirements is located in Figure 4.7.

	Length (a.a.)	aa's added to 5'	aa's added to 3'
LJFgene3	236 ǂ	-----	-----
LJFgene14	163	78	0
LJFgene1	225	14	0
LJFgene8	213	21	6
LJFgene9	130	80	31

ǂ Length does not account for gaps within the alignment

Figure 4.7. Number of amino acid residues added to each gene for multiple alignment.

When creating a codon alignment using the extended sequences, the corresponding nucleic acid sequences must match exactly. Because the amino acid sequences in the peptide alignment input field must correspond exactly to the nucleic acid sequences in the coding sequence input field, nucleotides were also added to the 5' and 3' ends of coding sequences (not the genomic sequences). The existence of

translational stop codons in the extended regions of the sequences interfered with the ability of the program to create the alignment. All intermittent stops in the extension regions of the peptide sequences were replaced with an R (letter representing an arginine residue) to extend the reading frame for acceptance by the PAL2NAL program. Arginine was chosen arbitrarily. The program indicated the position of the residue and codon sequence that did not correspond, thereby creating a record of the discrepancy (Figure 3.8).

A multiple alignment was also carried out using the nucleic acid sequences of gene family members as input, once with the coding sequence and again with the coding sequence plus enough nucleotides extending from the 5' and 3' end of each gene to make all sequences as long as the longest coding sequence (Figure 3.6). Both types of multiple alignment, those conducted using the peptide sequences and those conducted using nucleic acid sequences, were done to determine whether coding capacity once existed in those genes that appear to be non-functional or produce a truncated polypeptide. These alignments are located in Results Section 3.3.5. The alignments provide evidence that the sequence similarity between gene family members is isolated within the models on the 5' ends. The extension sequence of LJFgene14 used for these alignments is the natural extension of nucleotides of the coding reading frame. This input provides evidence that coding capacity does not exist beyond the gene border relative to LJFgene3. On the contrary, if the exact same alignment is created using a slightly altered sequence for LJFgene14, the coding capacity of the sequence upstream of the 5' start site of LJFgene14 remains intact. The sequence referred to uses two different reading frames. Within the model, the

sequence matching the open reading frame is used as input. The sequence used for the 5’ extension corresponds to a second reading frame that matches the original gene model. By combining these two segments into a single sequence, evidence exists that if some alteration of the sequence (the indel in the start codon and the cause of intron retention) had not occurred that lead to a reading frame shift, LJFgene14 and LJFgene3 would have nearly identical coding capacity.

Evidence for sequence similarity located downstream of the 3’ boundaries LJFgene3, LJFgene14, and LJFgene1 were analyzed using dot plot matrices. A comparison of the sequence beyond the 3’ borders of genes LJFgene3 and LJFgene1 reveals that similarity extends nearly 4 Kbp (Figure 3.13). The sequence similarity even goes so far that it extends into the most proximal 3’ neighbor gene on chromosome 3 (Figure 3.16). In the process of defining the boundaries of the coding capacity between models, a region of duplication was identified that contains two gene models on chromosome 3 and one model on chromosome 1. A blast search of the segment of correspondence from chromosome 3 against the genome does produce a match at the segment of chromosome 1 that contains the sequence downstream of LJFgene1. The algorithm predicts a gene putatively identified as a phosphotransferase on chromosome 3 within this sequence but does not predict a gene on chromosome 1. One plausible explanation is that there was a segmental rearrangement that moved this segment of LJFgene1 plus sequence extending into the neighbor gene into chromosome 1.

The family phylogeny was generated as an unrooted tree (Figure 3.5B) in order to reduce the likelihood of over interpretation of the relatedness of gene family

members. Though the broad plant family tree (Figure 3.5A) gives this family a root that indicates that LJFgene9 diverged first from the common ancestor of the entire family, followed by LJFgene8, then LJFgene14, LJFgene1, and LJFgene3 (in that order), this should not be interpreted as LJFgene9 being the "oldest" and LJFgene3 being the "youngest," or most recent, of the family. On the contrary, LJFgene3 is likely the gene that still most resembles the common ancestor to the entire family. Given the evidence for expression, it is hypothesized that this gene has been most highly conserved throughout its history to preserve function. With this logic in place, the phylogeny in Figure 3.5.B can be interpreted as LJFgene9 and LJFgene3 divergeing first from the common ancestor and the divergences of LJFgene14 and LJFgene1 from LJFgene3, as well as LJFgene8 from LJfgene9, came later, resulting in the synonymous substitution rates seen.

4.4. STRUCTURE, FUNCTION, AND LOCALIZATION PREDICTIONS

Numerous programs were used to attempt to predict the structure, function, and cellular location of the gene product for LJFgene3. Only the sequence for LJFgene3 was used as input for these programs since it is the most likely of the genes to produce a functional gene product.

Some of the results of these analyses did not agree. Results from CELLO (Table 3.11) indicate that the conceptually-translated protein sequence of LJFgene3 would most likely be localized to the nucleus or a chloroplast. The I-TASSER program predicted that the gene ontology (Results Section 3.4.3.3) cellular component of the same peptide sequence would most likely be cell periphery, i.e. extracellular. In

order to resolve this discrepancy, the reliability and predictive methods were evaluated. I-TASSER uses templates of known proteins from various species to make predictions. The template that best matches the query sequence in secondary and tertiary structure is used to infer function and location. All of the structural templates are β-lactamase molecules (Table 3.15), therefore the program predicts that the query peptide would act in a similar fashion to and be found in a similar location as a β-lactamase. CELLO is a program offered specifically for the identification of subcellular localization. It also compares an unknown sequence to peptides of known function, however the query sequence is first broken into many smaller sequences of equal length and each of these classified according to the subcellular location associated with amino acid composition. CELLO then uses a multi-level classification method based on support vector classifiers. In a comparison to other contemporary approaches, CELLO performed very well. In overall performance, (prediction of all localization possibilities) CELLO performed with 85% accuracy. However, in the prediction of each of the nuclear, plasma membrane, and extracellular locals, the accuracy score was over 90% [64]. Based on these methods of determination, the localization of the potential protein of LJFgene3 would perhaps be more accurately predicted by CELLO, and therefore be associated with some function in the nucleus or chloroplast. It is possible that both programs are correct. If the products of these genes were acting in a similar capacity to β-lactamase, for instance as an anti-fungal agent specific to plants, the possibility exists that their functional location could be intracellular as opposed to being secreted like β-lactamase in bacteria.

A Kyte-Doolittle hydropathy plot (Figure 3.18) was generated for LJFgene3 to rule out the possibility that the gene product is a transmembrane protein either on the plasma or nuclear membrane. The window size used was 19, making the standard criteria for hydrophobic regions to be any segment of line above a score of 1.6. Not a single line on the plot for LJFgene3 even reached 1.5, confirming that the potential peptide is not likely an integral membrane protein. As a model for comparison, a hydropathy plot for a known transmembrane protein, human rhodopsin, was included in the results. It appears to have 6, possibly 7, peaks that surpass the hydrophobic threshold and are composed of enough amino acid residues to span the plasma membrane. Human rhodopsin protein is known to have 7 alpha-helical transmembrane regions.

Both secondary structure prediction tools, PSI PRED and I-TASSER, agree with each other strongly and with high confidence (Figure 3.22). The amino terminus of the conceptually translated peptide of LJFgene3 appears to have 3 to 4 alpha-helices (depending on the program referenced), followed by four short beta sheets and terminating in a solitary carboxy-terminus alpha helix.

One tool for predicting tertiary structure and function, DomTHREADER, aligns the query sequence (conceptually translated peptide sequence of LJFgene3) and secondary structure with domains of known proteins linked to the PDB and CATH databases. Results for the pDomTHREADER output are located in Table 3.12 and Figure 3.24 in Section 3.4.5. All of the resulting domain matches were of low confidence and low sequence identity; however, a few of the matching domains had very similar secondary structure. The secondary structure of the domain with code

1v5sA00 exhibited a high degree of structural similarity to the carboxy-terminus of the query peptide sequence. The domain with code 1up8A00 was very similar to the animo-terminus and the domain with code 1m40A00 was most similar to the full length query than any other hits. The domain with code 1m40A00 has a class of alpha beta, architecture of a 3-layer (aba) sandwich, and the topology of a beta-lactamase. This is notable due to the results produced by the other predictive tools utilized for this study, I-TASSER.

I-TASSER uses known proteins as templates upon which unknown input sequences are compared based on a secondary structure alignment. The input sequence is threaded through a PDB library and subjected to multiple alignment algorithms. Profile-profile alignments have been shown to predict the correct topology templates even in cases where sequence identity with the query is low [73]. I-TASSER has been ranked number one for protein structure and function prediction in the community-wide Critical Assessment of Structure Prediction (CASP) experiments in recent years [74]. The results from this program for the query LJFgene3 conceptually translated peptide sequence all had low confidence scores and low identity; however it is noteworthy that a common theme persists through the results.

Aside from the threading templates (Table 3.14; only the top 3 results are beta-lactamase), all of the top structural analogs (Table 3.15), enzyme homologs (Table 3.16), and templates with similar binding sites (Table 3.18) are beta-lactamase molecules of bacterial species. Beta-lactamase proteins act as hydrolytic enzymes, cleaving a beta-lactam ring, and are a common mechanism of conferring bacterial resistance to classes of beta-lactam antibiotics including penicillin [80]. In gram-

negative bacteria, beta-lactamase molecules reside in the periplasm, the space beyond the plasma membrane. The protoenzyme is transported across the membrane in an unfolded state using the Sec apparatus or in a folded state using the Tat apparatus through recognition of a specific signal sequence that is cleaved post-translocation to produce the mature enzyme [81].

When comparing the highest scoring I-TASSER-generated structural model for gene LJFgene3 to the model of a known beta-lactamase molecule (Figure 3.26), it can be confirmed that similarity does exist. This does not imply that *Glycine max* produces an antibiotic beta lactamase. It simply suggests that a protein may be produced that has a similar structure and therefore possibly a similar function. Certain families of plants are known to produce antifungal and antibiotic molecules. For example the chitinase (*chi*) gene in beans and the ribosome-inactivating protein (*rip*) gene in barley have been transformed into soybean to produce a transgenic plant with multiple resistances [82]. Soybean is plagued by many pathogenic infections and naturally produces glyceollin, a phytoalexin known to have antibacterial, antinematodal, and antifungal activity. The glyceollin molecule is the end product of a pathway that adds a dimethylallyl to a pterocarpan precursor. The precursor molecule is (6aS, 11aS)-3,9,6a-trihydroxypterocarpan [(−)-glycinol] 4-dimethylallyltransferase, or G4DT. G4DT has been linked to a cDNA that contains a plastid-targeting signal and is presumed to be localized in the chloroplast. The fact that the glyceollin synthetic prenyltransferase is localized to the plastid further supports this presumption [83]. The second highest ranked possibility for the cellular location of LJFgene3 is in the chloroplast (Table 3.11). This detail, combined with the knowledge that the gene

product of LJFgene3 is presumed to have a similar function to the molecule produced by the biosynthetic pathway that utilizes G4DT, merits further research into any potential sequence similarity between genes.

An alignment of *Glycine max* G4DT mRNA-coding sequence and LJFgene3 coding sequence produced ambiguous results(data not included). A 500 bp difference exists between the two sequences. If only the aligned portion is considered, 316 out of 766, or 41% of, nucleotides are shared between the two. If the entire sequence of both is considered, only 316 out of 1234, or 25.6% of nucleotides shared. Given that the likelihood of a nucleotide at any one position in a sequence having a 1 in 4, or 25%, chance of being the same as the aligned sequence simply by random chance, these results are unimpressive. An alignment of amino acid residues further illustrated the differences between these genes and gene products. A BLAST of the G4DT coding sequence against the *Glycine max* genome did not produce any results matching a gene family member. The top hits were gene models with annotations linked to proteins of the prenyltransferase family. In addition, the tertiary structure (in the form of a ribbon diagram) of a pterocarpan (isoflavanone 4'-O-methyltransferase) from a related legume, *Medicago truncatula*, was accessed in PDB. The structure is displayed in Figure 4.8. A comparison of this structure with the predicted structural model of LJFgene3 from I-TASSER (Figure 3.25) reveals structural differences that likely indicate functional differences. It has not been ruled out that the gene product from this gene family may be involved in plant defense against pathogens, only that the genes are not part of the family of genes related to the glyceollin biosynthetic pathway.

Figure 4.8. Tertiary structure of isoflavanone 4'-O-methyltransferase from *Medicago truncatula* [84].

One of the primary aims of this research effort was to establish a plausible reason for the expansion of this family in soybean. The expansion could simply be a result of the soybean-specific WGD event predicted to have occurred around 13 Mya or could be tied to the function of the LJFgene family peptide. However, without first providing definitive evidence for gene product function, any attempts to address expansion would be purely speculative.

It has been taken into consideration that the gene product for this family may lack abundant expression evidence because expression is stress-induced (such as by a pathogen of environmental stressor) or tissue-specific. The EST library, found in Appendix F, contains information on the plant tissue from which each EST associated with this gene family was derived as well as the treatment that was applied to the

plant. Some tissues were exposed to drought stress or microorganisms prior to cDNA generation. Many of the tissues used came from seedlings, or young plants, but others came from specific parts such as root hairs. Table 3.9 contains some treatment data from the ESTs for this family. The question to be answered is: Can the tissue and treatment data reveal whether expression is tissue-specific or stress-induced? In order to perform a statistical test on the data, access to library information from the entire EST database for *Glycine max* is necessary. Without this data, a reliable test could not be performed. It is necessary to know what proportion of EST's for this species came from each cultivar, from a particular tissue, and using a certain treatment in order to determine the significance of the information found in the library for this gene family.

A second statistical test was considered for data collected for this gene family. PLACE data was compiled and organized by PLACE identifier and gene family member. The number of times a promoter element appears in the sequence upstream from the start site of a gene was recorded. Lastly, the data was organized according to associations with EST's. Those elements that appear only in genes that have EST data but not in genes without EST data, as well as elements appearing only in genes that lack EST data but not appearing in genes with EST data, were of particular interest. A statistician was consulted to determine if a statistical analysis could be applied to this data to determine whether the frequency of an element or the association of an element with only genes with a certain EST status could indicate significant expression patterns. Is the presence of an element necessary for transcription? Are elements associated with genes lacking EST data binding sites for a repressor? Unfortunately, no test could be devised due to the large quantity of data having such small sample

sizes. The raw data tables from which frequency data was calculated for cis-element identifiers can be found in Appendix G. The information reported in the Results Section is limited to identifiers, element sequence, and the genes in which they were found to be present in the promoter, and can be found in Table 3.8 in Results Section 3.4.2.

Each element in the PLACE database is distinguishable by an alphanumeric code, or identifier. This identifier is linked to descriptive data about the element such as whether it is required for expression of a particular gene or is a known binding site for a specific molecule, the organism from which the functional information was obtained, and citations for published research. Inspiration for possible deductions about expression patterns can be obtained from this information. Table 3.10 in Results Section 3.4.2 outlines patterns of overlapping occurrence of characteristics between cis-elements found in LJFgene family members. It is notable that in all five LJFgene family members, stress-induced elements exist in the promoter region such as various response elements (including dehydration-response), wound-induced, pathogen-induced, and light responsive elements. In addition, all five LJFgenes have elements for tissue-specific expression including a root/nodule element and elements found in seed or endosperm tissues. Of even more interest are those elements found to be present in the promoters of genes that have ESTs but absent in the genes that do not have ESTs. Two response elements (dehydration and sulfur) and one wound-induced element can be found in those genes with expression evidence. It is also noteworthy that a TATA box element found in the rice PAL gene promoter was detected in the three LJFgenes with expression evidence. The presence of these elements can be

associated with the treatments and tissues used for the soybean plants that produced ESTs for LJFgene3, LJFgene14, and LJFgene9. Treatments include pathogen-induced sensitivity and drought stress and the tissue source of 5 of the 13 ESTs were seeds, meristematic tissues, and root hairs with nodulating bacteria. It brings in to question whether any of these elements and treatments can be correlated to infer a source of differential expression or possible function. However, without a reliable statistical test, such inferences cannot be validated.

Transcription-dependent elements located in the promoter regions of the genes were researched but not all were added to the decorated models presented in Appendix E. Only those elements that were within a reasonably spaced window from the start site (or the 5' boundary of the first exon in the case of LJFgene14) were added to the decorated sequence. In addition, only those elements that were a reasonable distance from one another were added. In other words, TATA box and CAAT box elements that are separated by a distance that most closely correspond to the widely accepted data (approximately 100bp) have been added to the decorated sequences. It should be noted that only those sequences confirmed by the program were highlighted. These concensus sequences are not definitive. Variations from the consensus that are not highlighted but may be present in the necessary locations can still possess inherent affinity that will promote binding of the RNA polymerase subunits to the strand [15].

No polyadenylation signal sequences were highlighted in the decorated sequences. Although computer programs have been, and are being, developed to identify polyadenylation signals in plants, public access to these tools is not available and could not be utilized for the purpose of this study. Submission of a sequence

containing LJFgene3 to a prediction tool designed to identify polyA signals in human genes produced a record of the locations and sequence of signals known to exist in human genes. Only one signal matching the mammalian consensus polyA sequence was identified that is positioned within a reasonable distance of the stop codon. Without more evidence that cleavage occurs downstream of this site, its functional significance cannot be confirmed. Primary efforts to identify the polyA signal sequences in the LJFgene family were conducted through manual searches. Given the widely unpredictable nature of these signals in plants, a positive identification of their sequence and location could not be made with confidence. Although the signal is generally accepted to be an A-rich hexamer, a consensus sequence does not exist for plants as it does for metazoans making recognition by base sequence alone nearly impossible. The best hope of locating the signal was through EST analysis. The EST's for LJFgene3 do not agree on a 3' end making it difficult to predict a cleavage site. Without a definitive cleavage site, the information on accepted distances between the cleavage site and key elements cannot be applied.

4.5. NON-CODING SEQUENCE ANALYSIS

The non-coding sequences on chromosomes 19, 12, and 11 that resulted from the original BLAST searches were all hits for four of the five gene family members (Figure 3.1). They also displayed high conservation of sequence with each other as well as falling into a family motif (motif 2) as predicted by MEME (Figure 3.29). These criteria merited a deeper examination of those sequences.

Because the non-coding sequences were not recognized by the algortihms as being “genes,” no gene model was available in the genome browser and these sequences were referred to as “non-models” and earned the designation of “nm.”

Due to the strong conservation of sequence exhibited by these non-coding sequences (Figure 3.28) despite short sequence lengths, an investigation was conducted to determine whether they are microRNAs (miRNAs). High levels of conservation and dyad symmetry are two features of miRNAs. The conservation was evident, but the existence of dyad symmetry was more difficult to detect. In order to determine the patterns that would be produced from a sequence with dyad symmetry on a dot plot matrix, a series of tests involving the manipulation of a sequence known to have dyad symmetry was conducted. The result was that plotting a sequence with dyad symmetry against its reverse complement produced a distinguishable pattern. When the sequences for LJFnm19, LJFnm11, and LJFnm12 were plotted against their reverse complement, no patterns emerged. Despite lack of symmetry, the sequences were used as input in a miRNA prediction tool. The output (Figure 3.32) suggests that a short stretch of sequence in both LJFnm19 and LJFnm12 are precursory sequences for known miRNAs. This would provide a plausible explanation for the high degree of conservation exhibited by these sequences if the stretch of sequence representing the miRNA were the same between all three sequences, yet it was not the same sequence in LJFnm19 and LJFnm12, and a miRNA-linked sequence was not identified in LJFnm11.

The identity of the LJFnm sequences with LJFgene3 is very high, over 90% in all three non-coding sequences in the region of sequence similarity. This can be

viewed in the multiple alignment of LJFnm sequences against LJFgene3, LJFgene14, and LJFgene1 (Figure 3.30). The LJFnm sequences correspond to the sequence spanning the sixth exon, sixth intron, seventh exon, and the 3' UTR of the LJFgenes. Despite spanning two exons of the predicted family gene models (potentially 50 amino acids) and corresponding to conserved motifs, models were not predicted for the LJFnm sequences using FgenesH and GenomeScan. The question of why the algorithms did not predict a gene model for the non-coding sequences cannot be answered. Some critical element for gene prediction is likely missing.

The question is how did such a highly conserved sequence (nearly identical in composition and length) corresponding to a possible domain of a known transcribed gene end up on three different chromosomes? If patterns of neighboring genes could be established between LJFgenes and LJFnms, links could be attributed to duplication events. However, this is not the case. And if these sequences were relics of a gene that resulted from a duplication, it would be expected that at least some mutational differences would be apparent. If the proposed non-coding sequence had been isolated to a location on a single chromosome, rather than three, its existence could be attributed to a segmental duplication. This also is not the case. The possibility of the sequence being a transposable element can be ruled out due to a lack of inverted repeating sequences at the boundaries. Although it remains enigmatic, there is some important event from which this situation arose and more than likely something to be learned about this family by completing the puzzle.

4.6. FINAL CONCLUSIONS

The soybean LJFgene family contains five coding (LJFgene3, LJFgene1, LJFgene14, LJFgene8, and LJFgene9) and three non-coding sequences (LJFnm19, LJFnm11, and LJFnm12). Of the five coding sequences, three have EST data associated with them (LJFgene3, LJFgene14, and LJFgene9) and only one of these (LJFgene3) is presumed to produce a translated gene product, leaving the remaining genes to be considered pseudogenes. Data suggests that LJFgene3 is the most highly conserved (and therefore most representative of the original common ancestor of the family) and that the genes in the family diverged from this common ancestor in the following order: LJFgene9, LJFgene8, LJFgene14, and LJFgene1. Secondary structure and fold prediction programs have produced a possible structural model of the gene product; however, a high confidence potential function of the molecule could not be defined leaving the functional annotation as uncharacterized for now. The closest match in structure is a β-lactamase. The next step would be continued research into the possibility that this gene family has a role in soybean analogous to the role of β-lactamase in bacteria.

APPENDIX A.

GENE FAMILIES MEETING CHOICE CRITERIA

PFAM family	# genes via PFAM	PFAM functional annotation	*Glycine max* gene SupTab 5	Vvi:Ptr:Mtr:Gma:Ath:Aly:Cpa:Sbi:Zma:Bdi:Osa	Notes
PF06376	10	Protein of unknown function (DUF1070)	Glyma06g13060.1	0:4:1:8:2:2:2:0:0:0:0	still unknown function. Large intron region b/w 2 exons. 3 query matches: small section on chr 4 and 14.
PF06219	9	Protein of unknown function (DUF1005)	Glyma01g00910.1	1:0:1:3:1:1:1:1:1:1:1	Function still unknown. 10 hits in BLAST search.
PF03368	8	Domain of unknown function	Glyma13g22450.1	1:1:0:2:1:1:1:1:1:0:0	22 exons/21 introns Pfam:02170 PAZ domain Pfam:00035 Double-stranded RNA binding motif Pfam:00636 RNase3 domain Pfam:03368 Double stranded RNA binding domain Pfam:00270 DEAD/DEAH box helicase Pfam:00271 Helicase conserved C-terminal domain
PF04842	8	Plant protein of unknown function (DUF639)	Glyma19g29220.1	1:1:1:2:1:1:1:1:1:1:1	
PF05705	8	Eukaryotic protein of unknown function (DUF829)	Glyma01g37860.1	1:1:1:2:1:1:1:1:1:1:1	
PF07939	7	Protein of unknown function	Glyma07g34280.1	0:0:1:2:0:0:0:0:0:0:0	Function still unknown. 11 hits in BLAST.

		(DUF1685)			2 good 8 moderate matches
PF04949	6	Family of unknown function (DUF662)	Glyma19g37700.1	1:1:1:2:1:1:1:0:0:0:0	
PF06258	6	Protein of unknown function (DUF1022)	Glyma11g18750.1	2:3:0:5:2:2:2:2:2:2:2	11 exons:10 introns. Function unknown still. 6 BLAST matches--5 pretty strong
PF05811	6	Eukaryotic protein of unknown function (DUF842)	Glyma02g06480.1	1:1:1:5:3:3:1:1:1:1:1	Eukaryotic protein of unknown function (DUF842) Panther:21096 UNCHARACTERIZED KOG:3377Uncharacterized conserved protein
PF04134	5	Protein of unknown function, DUF393	Glyma13g44130.1	1:1:1:2:1:1:1:1:1:1:1	
PF06592	4	Protein of unknown function (DUF1138)	Glyma08g24930.1	0:1:0:4:1:1:0:0:0:0:0	function still unknown. 4 good BLAST hits
PF04502	4	Family of unknown function (DUF572)	Glyma13g33280.1	0:1:0:1:0:0:0:0:0:0:0	7introns:6exons. Family of unknown function, but CELL CYCLE CONTROL PROTEIN CWF16-RELATED in PANTHER. 7 BLAST hits-2strong
PF07386	4	Protein of unknown function (DUF1499)	Glyma08g39410.1	1:1:1:3:1:1:1:1:0:1:1	9 hits
PF04828	3	Protein of unknown function (DUF636)	Glyma11g33490.1, Glyma11g33500.1	0:1:0:1:0:0:0:0:0:0:0	
			Glyma11g33500.1		

PF05600	2	Protein of unknown function (DUF773)	Glyma02g30360.1	1:1:0:2:1:1:1:1:1:1:1	
PF06102	2	Domain of unknown function (DUF947)	Glyma03g02400.1	1:1:0:2:1:1:0:1:1:1:1	
PF04339	2	Protein of unknown function, DUF482	Glyma07g40290.1	1:1:0:2:1:1:1:1:1:1:1	
PF08538	2	Protein of unknown function (DUF1749)	Glyma17g02070.1	1:1:0:2:1:1:1:1:1:1:1	9exons:8introns protein of unknown function but KOG:4840 Predicted hydrolases or acyltransferases (alpha/beta hydrolase superfamily), 7 BLAST hits, 3 strong
PF04452	1	Protein of unknown function (DUF558)	Glyma20g27210.1	1:1:0:2:1:1:0:1:1:1:1	
PF06911	9	Senescence-associated protein	Glyma08g03770.1	1:4:1:4:0:0:1:1:1:1:1	senescence associated protein, family not names. 11 BLAST hits: 2 strong, 2 moderate.

APPENDIX B.

LJFgene FAMILY GENOMIC SEQUENCES (FASTA FORMAT)

>LJFgene3

TTTCGGTCTGTGAAGATATATGTCCATAAGTTCCTTAATTTTCTCGAACC
TTCATTTTCAGCTCCCAACAACAATGGCTTCAATGGCATCTTCAAGCTCC
TTCTGCAACCTCAAGTTCATCACCAAACCCAACAATGGTAGAAGAAGCTC
TCTTCCCCGTATTGTATTCTGTCAGAAGCACCACGATAGCACACCCACCG
ACCAAATCAACCGAAGGTTCTTATTTCTTCACACTCGCACTTTCTAATTC
CTTTCTATGGATTATTCATATCTATTCATACCCATCTTCTGAAATCTCTT
TATATTTCAATTATTTTGTCTATTGAAGAGAACTCATATTGAGAAGCAGC
GAAATAGCGACCATTGGTGCCATCTTGAACTTCGGGTACCCCTCCTCTGC
TTGTTTTTGGAAAATTTTTGTTTTTCATTTTATTTTGAATGTAAATTGAA
TTCAAGATTTGATTTTGTTGGTGGGTTTGAAGACCCTTTTGGTTTTTAAT
TTCGGTTTTGTTTTGTATTGGACATGGGTGGTGGTTAAAAAAGAGAAAAT
TGAGTTTGTGTCTTGTGTTTTGATGGTGCAGTGGGAAAAAACCTGATTAT
CTTGGAGTGCAGAAAAACCCACCAGCATTAGCTCTGTGCCCGGCAACGAA
GAATTGCGTGTCAACCTCTGAGAATATCAGTGATCGCACACATTATGCTC
CTCCATGGTAAAAGTTTCCTTCTTTTTCTTATTTTAATTTTCACCTTGGA
TTTATGGGATTATATGTATTAAATGCATTTTTTTAATTGTGTGTTTGGAC
AAACTACTTAGTTAAGTGCTCATCTCATCATGTAAGTGCTTATGCATAAG
TTGTTTCTATAATAAAAAAATAAAAATACATACGTATGAGTTGTTGTTGT
AAGCTTTTTTTCTTAAGTTATTCTGGAAATCTTATTGAAATAATCTGAAAA
CAACTTTTTTTTTTTACATGATCTGCTGAAAACAACTTATAGACATATGGT
AATCACATATCATTAAGTTAAGTTATTCCAAACACTTACATAAACACTTA
TAAGAGAAAAACAATAAGAAATAAAAACAAAATAAAATTTTCAATAAGTT
AAAATTAGTTTATAAAAGTTTTTTTTTATTATAGAAGCTCTCTTGTATTA
GCCTCTCCTAAAGTTTTTTTTTCTTCATAATTTAGCTTAAAAAGAAACCT
ATTTCATTTTTTCATTTTATTTTCTTCTCTTGTAAATGCTTTTGGAGAAG
TTTGTCCAAACATACCTTTACACAAATACTTATAAGATAAGTCTAATTAA
GCTTTTTCAAACACACTCAAAGTTAAAGTATTTCCATTTTGTTGTTTTTT
TGGGCCTTAATTTTCGGTTAACTAACTTGTGGTGTCTAAATTTCGTCTTG
AGGAGGGTCTGATAAACCAGATTTAGGAGATAAGTAAGCAGTCAGTGTTT
ATTTATTTATTTTTTGGATTTAGTCCAAAAGGAACCTATGGCATATTTGT
GAGAATCACGTTGCAATAGACAATAGTGCACCTGGAACGATTTATCACGT
TTTAAACAAGTAGTGCAACCGATATTGGACTATTGACTGTTGAACATTGT
TGACTTGACATAAGAATTGGACAATTGGTCACACACACATTGGCCGGTGA
AAGTAGTGCAATCTTTACTTTTTATTTCTTTTTGACAAAAAAATTTTCTC
TACCTTTGGCCCTTGTTTGAGCATGGTCGCCACAAAATCCAAACTTTATT
ATTGTATAGATGAATCATGATCTCACTTTGTTTTAGTATTTTCATTCTTT
TTCAGGAACTATAATCCTGAAGGTAGGAAAAAACCTGTGAACAGAGAGGA
AGCAATGGAGGAACTGATAGACGTGGTAATAAATCTAACTGAACTGAAAT
TTTGAATTATATCACTGGATTGCAATTTTCTTTTCCTTCCCTCTTTTAAT
CAACATCGATTATAATTTATAAAATTTATAAAAGGGGTGATAAACTGAAA
GTAAATATACACTTTGTAATGCACTATTCCTAACACACTTTCTATTATTC
ATTAAAATTTATTGAAAATTACGAAGTCATGGGTGGAATTCATTGAATAA
AGAGTAAGACCTACATGATTTTGTAATTTCTAATAAACTCTTACTATTAA
TAAAGAATGTATTTAAAAGGGTATGTTGTCAACATTTCTCTAATTTATAG
TTGATTTGTGATAATAGCTGGTTATTTGCACTTTTCCTTCCAATCATTGA
AGTAAAGTTAAACCAGTTCCGGACAATTTTGCAGATAGAATCAACAACAC

CAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGT
GTGGAGTACCAAAGCTCAATTTTGGGGGTAAGTGTAACTTACATCTAAGG
AAACTCATCACGAAGAAAAATGATAATTTTATACATTTGAGATGATATCA
AGATTCAAGAACCATCTCTAATTTCCTTCCCTCCTTTTTCTGTCATGTGC
TAACAGTTTGTAGATGATGTTGAGTTCTGGTTCCCACCGGGTAAGGGTTC
TACTGTGGAGTACCGATCTGCATCTCGGTTAGGAAACTTTGATTTTGATG
TGAACAGAAAAAGAATAAAGGTGTGATTTCATAATTCATGTGTTTTCTCT
ATAGTTAGATAAAGAAATTCTTGGTTCCATGGTAAAACTCCTCTTTCCTT
CATGTCATGTCAAACATTTTATACTCAAGTAGATGATTCACTAAATTTGA
GTCTCAAATGTTTTAACTTTATTCTAAATTAGTCACTTATTTTAACTGAA
GGTAAATTTGGTTAACTATGATCAGAAATACATTGACATTTTTTAATTGG
TAGAGATAAAGAATATTTTTTATGTACAATAAAGAGAGTATTTACTCCAG
AGGATGTAAATCCCTTGCTAAATATTTTTGTGATGAAAAATCTTGGTTGT
TGACAGGCACTGCGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCAAGA
CACCATATGATGAATAAACTCAGGCAGAATTAACATCAGCATCTAAGCAA
ATATTATTTCATATACTTTGTGACCTTGTATACATTTGTATTAGATACAA
ATCTCACAGGATCATTGAAAGCAAACTTTTCTTTGATTATTGGAATTGTA
GAGAAATCATTGAGAACAGTACTTCAAACTCTCGGGGAAGGAATGAAATG
AAGACCTTGCCCCATATCCTTCTTCAAGTTCATTAATTGGTCCGCTTATT
TACTCTTTCACCAAGTTCAATCTAACAATGTATCGTTCTTGTTTCAAGAA
TATTAATTTGTGTTTTGTTTCTAATGTGTTCTTCACGATTCACTTTTGTT
GTAAGTTTGACTTAGCTTCTTCATGTATTAAAACTTAACCTTGTGA

>LJFgene14
ACGTTCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAAC
AATAATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCTCAAGTTTA
TCACCAAACCCAACAATGGTAGAAGAAGCTCTCTTCGCCGTATTGTATTT
TGTCAGAAGCATCACGATGACACACCCACCGACCAAATCAACCGAAGGTT
CTTACTTCTTCACACTCACACTTTCTATTTCCTTTCTATTGATTATTCGT
AACCATCTTCTGAAATCTCGTTACATTTCAATTCTTTTGTGTATTGAAGA
GAACTCATATTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTCAA
CTTCGGGTACCCCTCCTCTGTTTTTGCTCTGTTTTTTTTTCTGGAAATTT
TAGTTTTTCATTTTATTTTGAATGTAAATTAAATTCGAGATTTGATTTTG
TTAGTGGGTGTTGAGACCCTTTTGGATTTTAGTTTGGGTTGTGTTTTGTA
TTGGAAATGGGTGGTTTGGGTTTTGTGTTTTGGTGGTGCAGTGGGAAAAA
ACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCATTAGCTCTGTGTC
CGCCAACTAAGAACTGCGTGTCAACCTCTGAGAATATCAGCGATCGCACA
CATTATGCTCCTCCATGGTAAAAGTTCCCTTCTTTTTCTTATTTTAATTT
TCACCTTGGATTTATGGGATTATATGAATTGAATGCAATTTTTTAATTGT
GTTTGGATAAACAACTTAGTTAACTGCCCATCATGTAAGTGCTTATGTAT
AAGGTTTTTCTATAGTAAAAAAATGTACAAGTTGCAAGCTTTTTTCTTAA
TTTATTCTTGAAATCTTATTGAAATAATCTGAGAACAACTTTTTCTTTTT
TACCTGTGATCTGAAAACAACTCATAGACATATCATAAACTTGGGATATA
GATAAATTTTTTCATAAACACTTACAAGAGAAAAAAAAATACAAAAGAAA
AATAAAATAAATTTTTCTATAAGCTAAAATTAGTGTATGTGTAAGCTAAT
TTGTAGTTCTTTCATATTAGCTTCTCCAAAAGTTTTTTTTTTTTTTAACT
TATGCATAAGTTAAATTTAGCTTAAAGATAAATTTATTTCATTTTCTTCT
CTTGTAAATGCTTTTGGAGAAATGTATCCAAACAGACCTTTACACAAGTA

CTAATAAGATAAGTCTAATTAAGCCTTTTCAAAACGCTCAAAGTTCAAGT
ATTTCCATTATGTTGATTCTTCGGGCCTTAATTTTCGGTTAAGTAACTTG
TGGTGTCAAAATTGCGTGTTGAGGAGGGTCTGGTAAACCAGATTTAGGAG
ATAAGTAATCAGGGTTTATTTATTTATTGGATTTAGTCCAAACGGAACCT
ATGGCCATATTTGTGAGAATCACGTTGCAATAATAGACAATGGTGCACTT
GGAACAATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAGCCGAAATT
GGACTATTGACTGTTGAACAATGTTGACTTGACATGAATTGGACTATTGG
TCACACACATTGGCCGGTGAAAGCAGTGCAATCTTAACTTTTTCTTTTTT
TTTTGACAACTTTTTTTTTTTCTCTATCTTTGGTCCTTGTATTGAGCATG
GTCCCACCAAAATCCAAACTTTATTATTGGACATAGATGAATCATGATGT
CACTTTGTTTTAATTTTTAGTTTCTTTTTCAGGAACTATAATCCTGAAGG
AAGGAAAAAACCTGTGAGCAGAGAGGAAGCAATGGAGGAACTGATAGACG
TGGTAATAAATCTAGCTGAACTGAAATCTTGAGTTATAACACTAGATTGC
AATTTTCTTTTCCTTCCCTCATTTTATCAACATCGATTATAATTTATAAA
ATTTATAAAAGGAGTGATAGTAAATATACACTTTGTAACACACTATTCTT
AACCCCACCCTTTTTATTATTGGTTAAAATTTATCAGAAATTAGAAAGTC
ATGAGTGGAACTCATTGAATAAAGAGTGAGACCTATGTGATTTTATAATT
TCTAATAAACTTATTATTAATAGTGTATTTAAAAGGATACATTGTGAACA
TTTCTCTAATTTATAATTGATTTGTAATAATAGCTGGTTATTTGCACTTT
TCCAATCATTGAAGTAAAGTTAAAACTAGTTCCGGATAATTTTACAGATC
GAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGA
AGACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGGTAAGTGTAA
CTTACATCTAAGGAAACTCATCTTGAAGAAAAATGATCATTTTATACATT
TGAGATGATATCAAGAACCATCTCTAATTTCCTTCTCCCTTTTTTCTGTC
ATGTGCTAACAGTTTGTGGATGATGTTGAGTTCTGGTTTCCACCCGGTAA
GGGTTCTACTGTGGAGTATCGATCTGCATCTCGGTTGGGAAACTTTGATT
TTGATGTGAACAGAAAAAGAATAAAGGTATGATTTCATAATTAATATGTG
CTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCAGGGTAAAACTCCC
CTTCCTTCATGTCATGTCAAACATTTTATACTTAAGTAGATTCACTAAAT
TTGAGTCTCAAATGTTTTAACTTTATTCTAAATTAGTCACTTATTTTAAC
GGAAGGTAAATTTGGTTGACTATGATGAGAAATACGTTGATATTTTTTAA
TTGGTAGAGATAAAGAATATTTTTTATGTACAATAAAGAGAGTATTTACT
CCAGAGGATGCAAATCCCTTACTAAATATTTTTGTGATGAAAAATCTTGG
TTGCTGACAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTC
AAGATACCATATGATTAATAAACTCAGGCTGAATTAGCATCAGCATCTAA
GCAAATATTATTTCATATACTTTGGACCTTGTATACTTTTGTATTAGATA
CAAATCGCACAGGATCATTGCAAGCAAACTTTTCTTAGATTTTTGGAATT
GTAGAGAAATCATTGAGAACAGTACTTCAAACTCTCTCGGGGAAGGAATG
AAATGAAGACCTTGCGCCATATCTTCTTCAAGTTCATTAATTGGTCCACT
TATTTACACCTTCACCGAGTTCAATCTAATAATGTATGAATCTTGTTTCA
AGAAATTTAATTTGTGTTTTGTTTCAAACGTGTTCTTCATGATTCACTTT
TGTTGTAAGTTTGACTTTGCTTCTTCATGTATTAAAGCTTATCCTTGCGA

>LJFgene1
CTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCT
CAAGTTTATCACCAAACCCAACAACAATGGTAGAACCAATGCTTCTTCTC
TTCCCCGTATTGTATTCTGTCAGAAGCACAACGATGACACCCCCACCGAC
CAAATCAACCGAAGGTTCTTACTTCTTCACACTCACACTCTCTCACTCCT

TCTTTCTATTGATTATTTATAACTATTCATACCCATCTTCTGAAATCTCT
TTACATTTCAATTCTTTTTGTGTATTGAAGAGAACTCATATTGAGAAGCA
GTGAAATAGCGACCATTGGTGCCATCTTCAACTTCGGGTACCCCTCCTCT
GTTTTTCCTCGGTTTTTTTCTTTTTTGGAAAATTTAGTTTTTTCATTTTA
TTTTGAATGTAAATTGTATTCAAGATTTGATTTTGTTGGTGGGTTTGGAG
ACCCTTTTGGATTTTAGTTTCAGTTTTGTATTGTATTGGAAATGGGTGGT
GGTTAAAAAAGAGAAAATTGAGTTTGGGTTTTGTGTTTTGGTGGTGCAGT
GGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCATTAGC
TCTGTGTCCGGCAACTAAGAACTGCGTGTCAACCTCTGAGAATATCAGTG
ATCGCACACATTATGCTCCTCCATGGTTAAAGTTCCCCTCTTTTTCTTAT
TTTAATTTTCACCTTCGATTTATGGGATTATATGAATTAAATGCAATTTT
TTTTACTTGTCTGTTTGGATAAACAACTTAGTTAAGTATTCATCATGTAA
GTGCTTATGTATAAGTTGTTTCTATAATAAATAAAAATGAGGAGGGAAAG
TTATTCCAAAAATCACTTTAAAAGAGGTACTCACTTTATTTTAATTATTG
ATTTTTTTAAAATTTAATGATTAAGATTAATTATTTATTATAATTATTAG
ATTCTAAAAAAATAAATAATAAGGATAATAAATAACTCTAAAAAAATTAT
TCTTAGAGCAAGTTGATGTTTTCCTAAAAAAAATATATAAATTGTTTATA
TAAGTCGTAAGCTTTTCGGAAATTATTCTTGAAATCTTATTGAAATAATC
TGAAAACAACTTTTTTTTACATGATTTGAAAACAACTTATAGACATATCA
TAATCACATATCATTAAGTTATTTTATTAAGTCATTTTCATAATTTATTT
CAAACACTTACATAAATACTTATAAGAGAAAATAAAATAAAATTAATTTC
TTTATAAGCTATAAAATTAGTTAATGTATAAGTCAATAGGTAGAAGCTCT
CTCGTATTAACTCTTCAAAAGTTTATTTTTTAACTTATATATAAGCTAAA
TTTAACTTAAAAGAGAAACTTATTTCATTTTTTCTCTTCCTTTCTTTTCT
AGTAAATGTTTTTATAAAAGTTTACCCAAACAGAACTTTACACAAGTACT
TATAAGATAAGTCTAATTAAGCTTTTTCAAACATGCTCAAAGTTAAAGTG
TTTCCATAATGTTTTTTTGGGCCTTAATTTTCGTCAAGTAACTTGTGATG
TCAAAATTGCGTGTTGAGGAGGGTCTGGTCAACTAGATTTAGGAGATAAT
TAAGCAGTCAGGGTTTATTTATTTATTGGATTTAATCCAAAAGGAACCTA
TGACATATTTGTGAGAATCACGTTGCAATAGATAATGGTGCACTTGGAAC
AATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAACCGATATTGGACT
ATTGACTGTTGAACATTGTTGACTTGACATAAGAAATGAACTATTGGTCA
CACACGCATTGGCCGGTGAAAGTAGTGCAATCTTTACTTTTTCTTTTTTT
GAAATTTTTTTTTTCACTACCTTTGGTCCTTGTATTGAGCATGGTCCCAC
CAAAATCCAAACTTTATTATTGGACATAGATGAATCATGATGTCACTTTG
TTTTAATATTTTCATTCTTTTTCAGGAACTATAATCCTGAAGGTAGGAAA
AAACCTGTGAGCAGGGAAGAAGCAATGGAGGAACTTATAGACGTGGTAAT
AAATGCAACTGAACTGAAATCTTGAGTTATCACTGGATTGAAATTTTCTT
TTCCTTCCCTCATTTTATCAACATTGATTATAATTTATAAAATTTATGAA
AGGAGTGATAGTGAATATACACTTTGTAACACTATTTCTAATACACTTTC
TATTATCGGTTAAAATTTATTGAAAACAGTGTTGTTAAATGGCGGCCATG
GCGGCGCCATGGCGGAGTTGCGTAACGGTTTTCTGAAAAAACGCCACCGA
ATAACGGTGGCGTGGCGGATTAAAGATGGCGGCGCCATGGCGGCCGCCAT
AGCCATGGCGGCCATGGCGGATGTGGCGGGGAGGCGGAAAATGGCAGAAT
TTTTTTTTTTGTCCGCGGTAGGAGTTGGGCTGACCCGATCCAACCCTACC
CGAAAACTTAATGAAAACCATGACCCCCCCTACCTTTCAGAACGCTGCTG
CAACCCTCGAAGCTTCAACATAGCGCGACCTCGGAACCAGCCACGACCCC
TGCAACCAGCCTCGGAACCAGCCGCGCTTGACCCGCTGCACACAGCAGCC

AGCAGCGACGACCCATGGCGTGAACGGCGGCGACGAGCACGGCGTGAACA
ACGGCGACGCGAACGGAGAAGACGACCCAGAACCGCGACCTCGACTTATA
CGGGTGGGTGGGCCAAAGCCTTTTTTTTGTTTTTGCTGCTTGACACCCCT
TTTTTATGTCTGCAGCTTTTTTTCCGTTTTTCTATTTGACACCCCCTTTT
ACTTTCGACAGTCCCATTTTTAATTTTTTTTTCTATTTGACACCACAATT
TTTTTTTCTGTTCAGTCCTCTTTTAAATGGCTGAACCATCATCCTCTTTA
ATGAGTTTATTTGGTGTTGGTACTTGATTATTGTATGAACTCATGAAACT
TTTTTAGTTTATTTGAATGCAATCCTTTGTTTTTTTCAATTTCAATGAGT
TTATATATATGTTTTTTTTTTTTTTTTGGTCCGCCATGACTTCCGCCATTT
TCCGCTACGCCATCCGCCATATTTTTATGGCGGATTTTTTACTTTCCGCC
ATGAACCGCCATCCGCCATTAACAACATTGATTGAAAATTATGAAGTCAT
GAGAGGAGCTCATTGAATAAAGAGTGAGAACTTACATGATTTTGTAATTT
CTAATAATTTTTTTACTATTAATAAAGAGTGTATTTAAATGGGTATGTTTT
TGAACATTTTTCTAATTTCTAATCGATTTGTAATAATAGCTGGTTATTTG
CACTTTCCCAATCATTGAAGTAAAGTTAAACCAGTTCCGGATAATTTTAC
AGATAGAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAGG
AAAGAAGACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGGTAAG
TGTAACTTACATCTAAGGAAACTCATCATGAAGAAAAATTATCCTTTTAT
ACATTTTAGATGATATCAAGATTCAAGAACCATCTTTAATTTCCTTCTCC
CTTTTTTCTGTCATGTGCTAACAGTTTGTGGATGATGTTGAGTTCTGGTT
TCCTCCGGGTAAGGGTTCTACTGTGGAGTATCGTTCTGCATCTCGGTTGG
GAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATGATTCCAT
AATTCATATGTGCTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCAG
GGTAAAACTCCCCTTTCCTTCATGTCATGTGAAGCATTTTTTACTCAAGT
AGATCCACTAAATTTGAGTCTCAAATGTTTTAACTTTATTCTAAATGTTT
TAACTTTATTTGAGTCTCAAATGTTTTAACTGAAGGTAAATTTGGTTAAC
TATGATCAGAAATACATTAACAAGAGTTGAAGAATATTTTTTATGTACAA
TAAAGAGAGTATTTGCTCGAGAGAATGTAAATCCTTTTCTAAATATTTTT
GTGATGAAAAATAATGGTTGCTGGCAGGCACTGAGACAAGAGTTGGAGAA
GAAAGGATGGGCATCTCAAGACACCATATGATGAAAAAACTTAGGCAGAA
TTCACATCAGCATCTAAGAAAATATTGTTTCATATACATTGTAACCTTGT
ATACTTTTGTATTAGATACAAAATCTCACAAGATCATTGAAAGCAAACTC
TTCATGATTATTGGAATTGTAGAAATGATTGAGAACAGTACTTCAAACTC
TCGGGGGAAGGAATGAAATGAAGATGTTACC

>LJFgene8
ATGGCTTCTTCGTTCTCCTTCTGCACCCTCAAGTTTCGCACCAAACCCAA
CGATAGTAGAAGCAGTGCTTCCTCTCTTCCCCGTATTCTATTCTGTCACA
ACCTCCACGATGACATTCACACACCCACTGACCAAATCAACCGAAGGTTC
ACTTCTCCACACTCTCATGCACTTTCTCACAGCTTTCTAATTTCTATTGA
TTATTCATAATTATTCATGCATACCCATCTTCTGAAATCTCTTTACACGT
CAATGATTTTGTATCTTAAAGACAACTCATATTGAGAAGCAGCGAAATAG
CGACCATCGGTGCCATCTTCGACTTCAGGTACCCCCGGCCCCTCCTCTGT
TTTTTAGAAATTTTTCTTTTCATTTTATTTTGAACGTAAATTTAATTCAA
GATTTGATTTTGTTAGTGAGTAGTGAGACCCTTTTGGTTTTTAGTATTAG
TTTTGTTTTGTATTGAAAATGGGTAGTTATTAAAAAAGAGGGTATTGTGT
GTTTTTTTTTTTTTTTTTTTTTTGTGGTGCAGTGGGAAAAAACCTGATTAT
CTTGGAGTGCAGAAAAACCCACCAGCTTTAGCTCTGTGTCCGGTAACTAG

GAACTGCGTATCAACCTCTGAGAATATCAGTGATCGCACTCATTATGCTC
CTCTTTGGTAAAAACTTTCCTTGTTTTCCTCATTTTAATTTTAGCCTTCC
TTTTCAAACCGCCAAGTTAATTTTTTAACCGTGTATTGGTGGTTTTTTTT
TGGTCCTTGAATGATACATAGGATTTTTCAAATAATTGGTTCAAAGACTA
ATTATATTTATAATATGTGATTGTAGTTGGTCAAAGGGGAAATTGTTTGA
TTTTTATCATGTATAAAATGTTTTTGGACCAATAATAATAACAAATAAAA
TAAGATTAATTTTTAACTCAATTGGTTGATTAAGAGTAAAAATATTTGTG
ATCTCTCATCAAAATAAAAAAATTTCTCTTCATTTTAAAACAAAAATCAC
ATCTTTAAAATTTAAAAGTCAGGTCCACACTTAAGATACTTAACTTAAAC
ATAGGTACTCAAACCAAAAATTTTCAAATATGTGAGATCCCACACAAATT
TTTATTTATTAGAAATCTTAAAATAAATTAAATGTTTATCAGGAATTTGA
ATGTATTTTAGCCTATTTTCGAACAAGGTTTTATCATACAAAATTTATGG
ATCCAACAAAATAAAAATAAGAAATTGTCATGGCTTGGTTTCCAAAACTG
AACCTCATGTTCAAGAACACGTGGAAGCACAACTGCAGCTTTGTCTTAGC
TCATCATCAAGCGCCAGTTAGCAATATTTTGCGGGAATAGACAATGAGAC
CAAAACCTTCATCATGCACAAGTCTATTTTAGTTGCACTAGGGTTTTGGA
GCCTTTCACAAGACCAAAACTAGATACGATTCATGCTAAAATAAAGCCTA
AAGCCTTTTGGAATTATTCACAGCAAAATCCGATGTTGACATCAACCATA
TGTAACATTATAATATATTACTAGCCTTTAAAATTCAATTCCACATTCAA
CAAAATTTATTACCCTGTTCTCTCTACACTACGCAATTAGAAAACAAAAG
GTTGAGCGGAGAAAAATCTAGAAAGTGCTTAGAAATTCAACTAAATTTTG
TCAAACAGGACGAGCCAATAAATACGATTAATGACAGCTACGAGAAATGT
ATATTTTTAGTTTAGAAACACGTCTTTAGTTATATCATAACAAAAAATAA
AAATTAAAGTTGTGTTTAAATATATGTTTTATTTAAAAAAATTAAAAATT
AGGAAATTGTGTTTTAATTCATGACTTTAAGTTAAAGTCACGAATTAAAA
GATGGCTTTAATAAAAGTAAATTTAAAAAATAAATAAAGAAGTCATGTTT
GAAAATATACCTTTCGTTGAAATGTTGTGCTTTTAAAACACAACTTGCTC
ATTTATTGATTTTTTAAAAATAAATTTACCTATTTTGGGATTTTTTTTTT
ATAAAAAATGCACCACTCTAGTAAGTAAACCTTTTTTTTTATATATAAAAA
ATTGCATTCAGGTAAACAACATAATTAAGTGTTTACTGATTCATTGAAAC
ACTTATGTATAAGTTGTTTATGTGATTGAAGAGAAAATAAAGTTAAATTA
TTTTCTTATAAATTGTAATATGTTTTCATGAGCTATGGAAAGTTTATTGA
AATAAACTGAAAATAGATTGTGGATATTTCATAAATACATATCTTAAAAT
TATTTTAAATATTCCAAACACTTATATAAATACTATAAGCACTTGTAATA
GAAGAAAGATAAGAATGTAAAATAAATTGATTTTTTTTCCATAAGTTGAT
TTAAGTTAAAATCAACTTATGTATCATAACACCTCAGGTTTTGAAAAAGT
TAAATGAGAGAGTTTTTAACAAAGTTAAGTGTATAAGTTAAATTTAAGAG
AAAAACACAATTTATTCTACCTTTTCTTTTCCCCATTTGTAATTGTTTAT
GGCCAATTTGATCCAAACAACATCTTTAGCCTAAATAAGCTCTTCCAATC
ACACTCTAAGTTTAAGTATGATTACTATGATGTTTAGATTTCATTGTGTT
GTTTTTTATGACCTTAATTTTCCGTCGAGTAACGCCATATTTGAATTAAA
GTTTGTAAATTTTAGATGAAATTCATTTTGCAAACTGAATTTGTAAAATC
AACTCTTCCTTTACTTCCAAAATCCAAACTTTATTATTGGACAAAGATGA
ATCGTGGAGTTACTTTGTTTTAAATTTTTAATTTTAATTATTGTTCAGGA
ACTACAATCCTGAAGGTAGGAAAAACCCTGTGAGCAGAGAAGAGGCAATG
GAGGAACTGATAGACGTGGTAATAAATCTAGCTGAAATCATAAGTTATTT
TCATGAATGCATTGCAATTTCCCTTTCCTAGCCTGTGTATCAACAAATGT
TATTATTTATAATAAGTTTAATTTTCATGCACTGACCGTATATAATAATT

TTATATTGATATCTAATCATAAATCATCATTTAAATTATTTTAAGATAAT
TAATTTAAAAGTTAATAAATTTACCGTATATAATGAATTATAATTGAATA
ATTGTATAAAAAATTTATAATGTCACTGTATAATTCTTTTTCTCATTTAT
AATTTGGTTGTAAGTTGTAGCTTGTAATAGCAGGTTATTTACCTTTTTCC
AATCATTGAAGTAAAGTTAAAATCCAGCTCTGGATAATTTTATAGATAGA
ATCAACAACACCAGACAAATTTACACCACGAATAGTTGAAAGGAAGGAAG
ACTATATTCATGTGGAGTACCAAAGCTCAATCTTGGGGGTATGTGTAACT
TACATCAAAAGGAAACTCATCGTGGAGAAAAATAATAATTTTGTACATTT
TAGATGATAATCAAGAACCATCTCTAATCCCTTCTCCCCTCCTTTTTATT
TTTTCTGCCATGTGCTAGCAGTTTGTGCATGATGTTGAGTTCTGGTTTCC
ACTGGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTCGGTTGGGGA
ACTTTGATTTTGATGTGAATAAGAAAAGAATAAAGGTATGTTTGTATCAT
TCCTTTGTGCTGTCTCGGTAG

>LJFgene9
AGCAACCATTGGTGCCATCTTCAACTTTAGGTACACTGCTTTATTGTTTT
CACAATGAGAATAGGTGACCAAAATTTAATTAATTCTTATAATATTTGAA
AATATTTCGAGAAATTTTATGTAAAATTAACTCTTTTTCTTTAAGATTTA
TGTGATTACTTCATAGCAAATTCTTGTCAAGTTTTAAGAACTTTTGAGTT
TTGATGTGTGTTTTCTATTGGAAATATGTGGGGATATATTATTTTGGATT
TTATGCTTCTTTGCAGAGGCAAAAAGCCAGATTATCTTGGAGTGCAGAAA
AATCAACCGGCATTAGCACTATGTCCGGCAACTAAGAACTGCATATCGAC
ATCTGAAAATGTCACTAACCTCACACATTACACTCCTCCTTGGTGAAATT
TCCTTCTTTATTTTTTTATTTATTAAAGTTTTAACTTTGGTTTATGATAT
TATCTGAATCTGAATTGGCTGCAAAGCCTGCAATTTGTTATTGAATATAT
ATTGTTGTTGAGTTTAATTTTACACTGAATAACTTTTAAAATAATTATTA
TAAAAATCAATAAAATTATTATCTATATAAAAAATATGATTAGATGATAA
AGTAAAACTTTTGTATGGTATTAATACATGAATTATTTTTTGTCCAGAAT
TTTTCAGATTGTCATAAATTTGCATTTTGATAAAGTGTATAATTCACTCA
ATTTCAAAATAACTGACTTTTTAACACCAATTTGGCAAAATGTTGCTTAT
GGAACAATTTAATTAATTATCACATTTTAAAACCATTGAAAGTTCAATTG
GTGTTGCCGTGTTGGGCGCATTGAGCCCTTAACTTAGGAAATGTATTATC
TCATGTTGAATTTGCTTTTTTTTTTTTTTTTTATTTCTCTTCAGCTAAGTT
CCGAGCATTGAGCCTTTAACTTAGGAAATGTATTATCACATGTTGGGTAT
AGATTATTAATTAAGAGAACATTGATTATTCTACAGGAACTACAATCCTG
AAGGTAGGAAAGATCATGTGAGCAAAGAGGCAATGGAGGAACTGATAGAT
GTGGTAATTTTAATTAGGATCATGTTAAGTCTTAAGCTACTTAGTTAAAG
AATCAATAAACAATTTTTTTAGGAAATCACGTGATTAATTACATTAGAAA
TCACAGTTACACTGGCAATACTGTTATATAAAAACTATTTATTTTAAATT
GGTTGACTAGTATCTTAAGAACACTTATTAAAAAAGTGTATTGAGTTTTT
GTATCTCAATAGGAGTATAATTAAGACTATTTATAGTTGGTTTGTTATAG
CTACTGATGAGTTTTCAAATCATTAAAGTTACAAAAATCAGTTACTATTG
TTTTTTATTTTACAGATAGAATCGACAATACTACCAGAAAATTTTACACC
AAGGATTGTAGAAAGAACAGAAGATTATCTTAGATTGGAATACCAAAGTG
TATACAAGCCACAAATTTTAACTTCAATGTCACCAATATCATTGTATGCA
GAAAAAATGAATAGTAACTTTTTACTATTAGACTGAAAA

APPENDIX C.

LJFgene FAMILY CODING SEQUENCES (FASTA FORMAT)

>LJFgene3
ATGTCCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAAC
AACAATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCTCAAGTTCA
TCACCAAACCCAACAATGGTAGAAGAAGCTCTCTTCCCCGTATTGTATTC
TGTCAGAAGCACCACGATAGCACACCCACCGACCAAATCAACCGAAGAGA
ACTCATATTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTGAACT
TCGGTGGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCA
TTAGCTCTGTGCCCGGCAACGAAGAATTGCGTGTCAACCTCTGAGAATAT
CAGTGATCGCACACATTATGCTCCTCCATGGAACTATAATCCTGAAGGTA
GGAAAAAACCTGTGAACAGAGAGGAAGCAATGGAGGAACTGATAGACGTG
ATAGAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAGGAA
AGAAGACTATATTCGTGTGGAGTACCAAAGCTCAATTTTGGGGTTTGTAG
ATGATGTTGAGTTCTGGTTCCCACCGGGTAAGGGTTCTACTGTGGAGTAC
CGATCTGCATCTCGGTTAGGAAACTTTGATTTTGATGTGAACAGAAAAAG
AATAAAGGCACTGCGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCAAG
ACACCATA

>LJFgene14
ATGGGTGGTTTGGGTTTTGTGTTTTGGTGGTGCAGTGGGAAAAAACCTGA
TTATCTTGGAGTGCAGAAAAACCCACCAGCATTAGCTCTGTGTCCGCCAA
CTAAGAACTGCGTGTCAACCTCTGAGAATATCAGCGATCGCACACATTAT
GCTCCTCCATGGAACTATAATCCTGAAGGAAGGAAAAAACCTGTGAGCAG
AGAGGAAGCAATGGAGGAACTGATAGACGTGATCGAATCAACAACACCAG
ACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTG
GAGTACCAAAGCTCAATCTTGGGGTTTGTGGATGATGTTGAGTTCTGGTT
TCCACCCGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTCGGTTGG
GAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGCACTGAGACAA
GAGTTGGAGAAGAAAGGATGGACATCTCAAGATACCATA

>LJFgene1
ATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCTCAAGTTTATCAC
CAAACCCAACAACAATGGTAGAACCAATGCTTCTTCTCTTCCCCGTATTG
TATTCTGTCAGAAGCACAACGATGACACCCCCACCGACCAAATCAACCGA
AGAGAACTCATATTGAGAAGCAGTGAAATAGCGACCATTGGTGCCATCTT
CAACTTCGGTGGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCAC
CAGCATTAGCTCTGTGTCCGGCAACTAAGAACTGCGTGTCAACCTCTGAG
AATATCAGTGATCGCACACATTATGCTCCTCCATGGAACTATAATCCTGA
AGGTAGGAAAAAACCTGTGAGCAGGGAAGAAGCAATGGAGGAACTTATAG
ACGTGATAGAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAA
AGGAAAGAAGACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGTT
TGTGGATGATGTTGAGTTCTGGTTTCCTCCGGGTAAGGGTTCTACTGTGG
AGTATCGTTCTGCATCTCGGTTGGGAAACTTTGATTTTGATGTGAACAGA
AAAAGAATAAAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGGCATC
TCAAGACACCATA

>LJFgene8
ATGGCTTCTTCGTTCTCCTTCTGCACCCTCAAGTTTCGCACCAAACCCAA
CGATAGTAGAAGCAGTGCTTCCTCTCTTCCCCGTATTCTATTCTGTCACA

ACCTCCACGATGACATTCACACACCCACTGACCAAATCAACCGAAGACAA
CTCATATTGAGAAGCAGCGAAATAGCGACCATCGGTGCCATCTTCGACTT
CAGTGGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCTT
TAGCTCTGTGTCCGGTAACTAGGAACTGCGTATCAACCTCTGAGAATATC
AGTGATCGCACTCATTATGCTCCTCTTTGGAACTACAATCCTGAAGGTAG
GAAAAACCCTGTGAGCAGAGAAGAGGCAATGGAGGAACTGATAGACGTGA
TAGAATCAACAACACCAGACAAATTTACACCACGAATAGTTGAAAGGAAG
GAAGACTATATTCATGTGGAGTACCAAAGCTCAATCTTGGGGTTTGTGCA
TGATGTTGAGTTCTGGTTTCCACTGGGTAAGGGTTCTACTGTGGAGTATC
GATCTGCATCTCGGTTGGGGAACTTTGATTTTGATGTGAATAAGAAAAGA
ATAAAGGTATGTTTGTATCATTCCTTTGTGCTGTCTCGG

>LJFgene9
GCAACCATTGGTGCCATCTTCAACTTTAGAGGCAAAAAGCCAGATTATCT
TGGAGTGCAGAAAAATCAACCGGCATTAGCACTATGTCCGGCAACTAAGA
ACTGCATATCGACATCTGAAAATGTCACTAACCTCACACATTACACTCCT
CCTTGGAACTACAATCCTGAAGGTAGGAAAGATCATGTGAGCAAAGAGGC
AATGGAGGAACTGATAGATGTGATAGAATCGACAATACTACCAGAAAATT
TTACACCAAGGATTGTAGAAAGAACAGAAGATTATCTTAGATTGGAATAC
CAAAGTGTATACAAGCCACAAATTTTAACTTCAATGTCACCAATATCATT
GTATGCAGAAAAAATGAATAGTAACTTTTTACTATTAGAC

APPENDIX D.

LJFgene FAMILY CONCEPTUALLY TRANSLATED PEPTIDE SEQUENCES (FASTA FORMAT)

>LJFgene3
MSISSLIFSNLHFQLPTTMASMASSSSFCNLKFITKPNNGRRSSLPRIVF
CQKHHDSTPTDQINRRELILRSSEIATIGAILNFGGKKPDYLGVQKNPPA
LALCPATKNCVSTSENISDRTHYAPPWNYNPEGRKKPVNREEAMEELIDV
IESTTPDKFSPRIVERKEDYIRVEYQSSILGFVDDVEFWFPPGKGSTVEY
RSASRLGNFDFDVNRKRIKALRQELEKKGWASQDTI

>LJFgene14
MGGLGFVFWWCSGKKPDYLGVQKNPPALALCPPTKNCVSTSENISDRTHY
APPWNYNPEGRKKPVSREEAMEELIDVIESTTPDKFSPRIVERKEDYIRV
EYQSSILGFVDDVEFWFPPGKGSTVEYRSASRLGNFDFDVNRKRIKALRQ
ELEKKGWTSQDTI

>LJFgene1
MASMASSSSFCNLKFITKPNNNGRTNASSLPRIVFCQKHNDDTPTDQINR
RELILRSSEIATIGAIFNFGGKKPDYLGVQKNPPALALCPATKNCVSTSE
NISDRTHYAPPWNYNPEGRKKPVSREEAMEELIDVIESTTPDKFSPRIVE
RKEDYIRVEYQSSILGFVDDVEFWFPPGKGSTVEYRSASRLGNFDFDVNR
KRIKALRQELEKKGWASQDTI

>LJFgene8
MASSFSFCTLKFRTKPNDSRSSASSLPRILFCHNLHDDIHTPTDQINRRQ
LILRSSEIATIGAIFDFSGKKPDYLGVQKNPPALALCPVTRNCVSTSENI
SDRTHYAPLWNYNPEGRKNPVSREEAMEELIDVIESTTPDKFTPRIVERK
EDYIHVEYQSSILGFVHDVEFWFPLGKGSTVEYRSASRLGNFDFDVNKKR
IKVCLYHSFVLSR

>LJFgene9
ATIGAIFNFRGKKPDYLGVQKNQPALALCPATKNCISTSENVTNLTHYTP
PWNYNPEGRKDHVSKEAMEELIDVIESTILPENFTPRIVERTEDYLRLEY
QSVYKPQILTSMSPISLYAEKMNSNFLLLD

APPENDIX E.

LJFgene FAMILY MEMBER DECORATED SEQUENCES

Start codon Stop codon exon intron promoter element

LJFgene3

ATCTTCTTTGACCAATGACCACAAAATACTTTTAGTCAAATAAGTGATTT
GGAGACAACCATAATTTCTAATTCAAGAAAATAAAATCTAAAATGAAACT
TATTCTAAAAAGTAAACGGATATAATAACATCCTTCACAAAAACCTTTTC
GTGCATATAAGCATTTCTGGATATTTGTATTATCTGAAATTGGAAATAGT
CCACTATCTTAAAAACATCGAAATATAATTTTTTTTCCAAATATCCAAAT
CCATAAAAAAGAAAAAAAAAAAGTCCCACCGCCCACCTTCTTTATCACAT
GATTCACATCTCATTCCTTATATTTGGTTCACATTCTTAAATTATAAATA
TTTCGGTCTGTGAAGATATATGTCCATAAGTTCCTTAATTTTCTCGAACC
TTCATTTTCAGCTCCCAACAACAATGGCTTCAATGGCATCTTCAAGCTCC
TTCTGCAACCTCAAGTTCATCACCAAACCCAACAATGGTAGAAGAAGCTC
TCTTCCCCGTATTGTATTCTGTCAGAAGCACCACGATAGCACACCCACCG
ACCAAATCAACCGAAGGTTCTTATTTCTTCACACTCGCACTTTCTAATTC
CTTTCTATGGATTATTCATATCTATTCATACCCATCTTCTGAAATCTCTT
TATATTTCAATTATTTTGTCTATTGAAGAGAACTCATATTGAGAAGCAGC
GAAATAGCGACCATTGGTGCCATCTTGAACTTCGGGTACCCCTCCTCTGC
TTGTTTTTGGAAAATTTTTGTTTTTCATTTTATTTTGAATGTAAATTGAA
TTCAAGATTTGATTTTGTTGGTGGGTTTGAAGACCCTTTTGGTTTTTAAT
TTCGGTTTTGTTTTGTATTGGACATGGGTGGTGGTTAAAAAAGAGAAAAT
TGAGTTTGTGTCTTGTGTTTTGATGGTGCAGTGGGAAAAAACCTGATTAT
CTTGGAGTGCAGAAAAACCCACCAGCATTAGCTCTGTGCCCGGCAACGAA
GAATTGCGTGTCAACCTCTGAGAATATCAGTGATCGCACACATTATGCTC
CTCCATGGTAAAAGTTTCCTTCTTTTTCTTATTTTAATTTTCACCTTGGA
TTTATGGGATTATATGTATTAAATGCATTTTTTTAATTGTGTGTTTGGAC
AAACTACTTAGTTAAGTGCTCATCTCATCATGTAAGTGCTTATGCATAAG
TTGTTTCTATAATAAAAAAATAAAAATACATACGTATGAGTTGTTGTTGT
AAGCTTTTTTCTTAAGTTATTCTGGAAATCTTATTGAAATAATCTGAAAA
CAACTTTTTTTTTTACATGATCTGCTGAAAACAACTTATAGACATATGGT
AATCACATATCATTAAGTTAAGTTATTCCAAACACTTACATAAACACTTA
TAAGAGAAAAACAATAAGAAATAAAAACAAAATAAAATTTTCAATAAGTT
AAAATTAGTTTATAAAAGTTTTTTTTTATTATAGAAGCTCTCTTGTATTA
GCCTCTCCTAAAGTTTTTTTTTCTTCATAATTTAGCTTAAAAAGAAACCT
ATTTCATTTTTTCATTTTATTTTCTTCTCTTGTAAATGCTTTTGGAGAAG
TTTGTCCAAACATACCTTTACACAAATACTTATAAGATAAGTCTAATTAA
GCTTTTTCAAACACACTCAAAGTTAAAGTATTTCCATTTTGTTGTTTTTT
TGGGCCTTAATTTTCGGTTAACTAACTTGTGGTGTCTAAATTTCGTCTTG
AGGAGGGTCTGATAAACCAGATTTAGGAGATAAGTAAGCAGTCAGTGTTT
ATTTATTTATTTTTTGGATTTAGTCCAAAAGGAACCTATGGCATATTTGT
GAGAATCACGTTGCAATAGACAATAGTGCACCTGGAACGATTTATCACGT
TTTAAACAAGTAGTGCAACCGATATTGGACTATTGACTGTTGAACATTGT
TGACTTGACATAAGAATTGGACAATTGGTCACACACACATTGGCCGGTGA
AAGTAGTGCAATCTTTACTTTTTATTTCTTTTTGACAAAAAAATTTTCTC
TACCTTTGGCCCTTGTTTGAGCATGGTCGCCACAAAATCCAAACTTTATT
ATTGTATAGATGAATCATGATCTCACTTTGTTTTAGTATTTTCATTCTTT
TTCAGGAACTATAATCCTGAAGGTAGGAAAAAACCTGTGAACAGAGAGGA
AGCAATGGAGGAACTGATAGACGTGGTAATAAATCTAACTGAACTGAAAT
TTTGAATTATATCACTGGATTGCAATTTTCTTTTCCTTCCCTCTTTTAAT
CAACATCGATTATAATTTATAAAATTTATAAAAGGGGTGATAAACTGAAA
GTAAATATACACTTTGTAATGCACTATTCCTAACACACTTTCTATTATTC
ATTAAAATTTATTGAAAATTACGAAGTCATGGGTGGAATTCATTGAATAA

AGAGTAAGACCTACATGATTTTGTAATTTCTAATAAACTCTTACTATTAA
TAAAGAATGTATTTAAAAGGGTATGTTGTCAACATTTCTCTAATTTATAG
TTGATTTGTGATAATAGCTGGTTATTTGCACTTTTCCTTCCAATCATTGA
AGTAAAGTTAAACCAGTTCCGGACAATTTTGCAGATAGAATCAACAACAC
CAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGT
GTGGAGTACCAAAGCTCAATTTTGGGGGTAAGTGTAACTTACATCTAAGG
AAACTCATCACGAAGAAAAATGATAATTTTATACATTTGAGATGATATCA
AGATTCAAGAACCATCTCTAATTTCCTTCCCTCCTTTTTCTGTCATGTGC
TAACAGTTTGTAGATGATGTTGAGTTCTGGTTCCCACCGGGTAAGGGTTC
TACTGTGGAGTACCGATCTGCATCTCGGTTAGGAAACTTTGATTTTGATG
TGAACAGAAAAAGAATAAAGGTGTGATTTCATAATTCATGTGTTTTCTCT
ATAGTTAGATAAAGAAATTCTTGGTTCCATGGTAAAACTCCTCTTTCCTT
CATGTCATGTCAAACATTTTATACTCAAGTAGATGATTCACTAAATTTGA
GTCTCAAATGTTTTAACTTTATTCTAAATTAGTCACTTATTTTAACTGAA
GGTAAATTTGGTTAACTATGATCAGAAATACATTGACATTTTTTAATTGG
TAGAGATAAAGAATATTTTTTATGTACAATAAAGAGAGTATTTACTCCAG
AGGATGTAAATCCCTTGCTAAATATTTTTGTGATGAAAAATCTTGGTTGT
TGACAGGCACTGCGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCAAGA
CACCATATGATGAATAAACTCAGGCAGAATTAACATCAGCATCTAAGCAA
ATATTATTTCATATACTTTGTGACCTTGTATACATTTGTATTAGATACAA
ATCTCACAGGATCATTGAAAGCAAACTTTTCTTTGATTATTGGAATTGTA
GAGAAATCATTGAGAACAGTACTTCAAACTCTCGGGGAAGGAATGAAATG
AAGACCTTGCCCCATATCCTTCTTCAAGTTCATTAATTGGTCCGCTTATT

LJFgene14

AACCTTTTCGTACAGATAAGCATTTATGGATATTTTTAGTATCCAAAATT
GTCACTTTCTCAAACAATCGAAATATATACTATTTATTTTCTAAATATCT
AAATCCATAAAAAGGAAAATAAATAAAAAAATAAAAATGTTGCGGAAACG
AAGTCCCACCTTCTTTTATTCATCACATGATTCACATCTCATTTCTTATT
TTCGGGTCACTTTGTTAAATTATAAATAATTTCGTTCTGTGAAGGTACAC
ACGTTCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAAC
AATAATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCTCAAGTTTA
TCACCAAACCCAACAATGGTAGAAGAAGCTCTCTTCGCCGTATTGTATTT
TGTCAGAAGCATCACGATGACACACCCACCGACCAAATCAACCGAAGGTT
CTTACTTCTTCACACTCACACTTTCTATTTCCTTTCTATTGATTATTCGT
AACCATCTTCTGAAATCTCGTTACATTTCAATTCTTTTGTGTATTGAAGA
GAACTCATATTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTCAA
CTTCGGGTACCCCTCCTCTGTTTTTGCTCTGTTTTTTTTTCTGGAAATTT
TAGTTTTTCATTTTATTTTGAATGTAAATTAAATTCGAGATTTGATTTTG
TTAGTGGGTGTTGAGACCCTTTTGGATTTTAGTTTGGGTTGTGTTTTGTA
TTGGAAATGGGTGGTTTGGGTTTTGTGTTTTGGTGGTGCAGTGGGAAAAA
ACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCATTAGCTCTGTGTC
CGCCAACTAAGAACTGCGTGTCAACCTCTGAGAATATCAGCGATCGCACA
CATTATGCTCCTCCATGGTAAAAGTTCCCTTCTTTTTCTTATTTTAATTT
TCACCTTGGATTTATGGGATTATATGAATTGAATGCAATTTTTTAATTGT
GTTTGGATAAACAACTTAGTTAACTGCCCATCATGTAAGTGCTTATGTAT
AAGGTTTTTCTATAGTAAAAAAATGTACAAGTTGCAAGCTTTTTTCTTAA
TTTATTCTTGAAATCTTATTGAAATAATCTGAGAACAACTTTTTCTTTTT

TACCTGTGATCTGAAAACAACTCATAGACATATCATAAACTTGGGATATA
GATAAATTTTTTCATAAACACTTACAAGAGAAAAAAAAATACAAAAGAAA
AATAAAATAAATTTTTCTATAAGCTAAAATTAGTGTATGTGTAAGCTAAT
TTGTAGTTCTTTCATATTAGCTTCTCCAAAAGTTTTTTTTTTTTTTAACT
TATGCATAAGTTAAATTTAGCTTAAAGATAAATTTATTTCATTTTCTTCT
CTTGTAAATGCTTTTGGAGAAATGTATCCAAACAGACCTTTACACAAGTA
CTAATAAGATAAGTCTAATTAAGCCTTTTCAAAACGCTCAAAGTTCAAGT
ATTTCCATTATGTTGATTCTTCGGGCCTTAATTTTCGGTTAAGTAACTTG
TGGTGTCAAAATTGCGTGTTGAGGAGGGTCTGGTAAACCAGATTTAGGAG
ATAAGTAATCAGGGTTTATTTATTTATTGGATTTAGTCCAAACGGAACCT
ATGGCCATATTTGTGAGAATCACGTTGCAATAATAGACAATGGTGCACTT
GGAACAATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAGCCGAAATT
GGACTATTGACTGTTGAACAATGTTGACTTGACATGAATTGGACTATTGG
TCACACACATTGGCCGGTGAAAGCAGTGCAATCTTAACTTTTTCTTTTTT
TTTTGACAACTTTTTTTTTTTCTCTATCTTTGGTCCTTGTATTGAGCATG
GTCCCACCAAAATCCAAACTTTATTATTGGACATAGATGAATCATGATGT
CACTTTGTTTTAATTTTTAGTTTCTTTTTCAGGAACTATAATCCTGAAGG
AAGGAAAAAACCTGTGAGCAGAGAGGAAGCAATGGAGGAACTGATAGACG
TGGTAATAAATCTAGCTGAACTGAAATCTTGAGTTATAACACTAGATTGC
AATTTTCTTTTCCTTCCCTCATTTTATCAACATCGATTATAATTTATAAA
ATTTATAAAAGGAGTGATAGTAAATATACACTTTGTAACACACTATTCTT
AACCCCACCCTTTTTATTATTGGTTAAAATTTATCAGAAATTAGAAAGTC
ATGAGTGGAACTCATTGAATAAAGAGTGAGACCTATGTGATTTTATAATT
TCTAATAAACTTATTATTAATAGTGTATTTAAAAGGATACATTGTGAACA
TTTCTCTAATTTATAATTGATTTGTAATAATAGCTGGTTATTTGCACTTT
TCCAATCATTGAAGTAAAGTTAAAACTAGTTCCGGATAATTTTACAGATC
GAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGA
AGACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGGTAAGTGTAA
CTTACATCTAAGGAAACTCATCTTGAAGAAAAATGATCATTTTATACATT
TGAGATGATATCAAGAACCATCTCTAATTTCCTTCTCCCTTTTTTCTGTC
ATGTGCTAACAGTTTGTGGATGATGTTGAGTTCTGGTTTCCACCCGGTAA
GGGTTCTACTGTGGAGTATCGATCTGCATCTCGGTTGGGAAACTTTGATT
TTGATGTGAACAGAAAAAGAATAAAGGTATGATTTCATAATTAATATGTG
CTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCAGGGTAAAACTCCC
CTTCCTTCATGTCATGTCAAACATTTTATACTTAAGTAGATTCACTAAAT
TTGAGTCTCAAATGTTTTAACTTTATTCTAAATTAGTCACTTATTTTAAC
GGAAGGTAAATTTGGTTGACTATGATGAGAAATACGTTGATATTTTTTAA
TTGGTAGAGATAAAGAATATTTTTTATGTACAATAAAGAGAGTATTTACT
CCAGAGGATGCAAATCCCTTACTAAATATTTTTGTGATGAAAAATCTTGG
TTGCTGACAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTC
AAGATACCATATGATTAATAAACTCAGGCTGAATTAGCATCAGCATCTAA
GCAAATATTATTTCATATACTTTGGACCTTGTATACTTTTGTATTAGATA
CAAATCGCACAGGATCATTGCAAGCAAACTTTTCTTAGATTTTTGGAATT
GTAGAGAAATCATTGAGAACAGTACTTCAAACTCTCTCGGGGAAGGAATG
AAATGAAGACCTTGCGCCATATCTTCTTCAAGTTCATTAATTGGTCCACT

LJFgene1

AATTAAAAAATGTAAGCATTATTATTTTTACCAATTTGACAAAAACCTTT
TCGTACAGAGATAAGCATTTGTGGATATTTTTAGTATCCAAAATTGTCCA
TTTTCTAAAAAAATCGAAATATATATTTATTTTCTAAATATCCAAATCCA
TAAAAAGGAGAAAGAAAAAACAAAAAAAAAATGTTGCGGAAACGAAGCGT
CCCACCACCCACCTTCTTTTTATATTCATCATTACATGATTCACATCTCA
TTCCATATTTTCGGGTCACATTCTCAAATTATTATAACTAATTTCGTCAT
GTGAAGATACGTTCATAAGTTCCTTAATTTTCTTGAACCTTTATTTTCAG
CTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCT
CAAGTTTATCACCAAACCCAACAACAATGGTAGAACCAATGCTTCTTCTC
TTCCCCGTATTGTATTCTGTCAGAAGCACAACGATGACACCCCCACCGAC
CAAATCAACCGAAGGTTCTTACTTCTTCACACTCACACTCTCTCACTCCT
TCTTTCTATTGATTATTTATAACTATTCATACCCATCTTCTGAAATCTCT
TTACATTTCAATTCTTTTTGTGTATTGAAGAGAACTCATATTGAGAAGCA
GTGAAATAGCGACCATTGGTGCCATCTTCAACTTCGGGTACCCCTCCTCT
GTTTTTCCTCGGTTTTTTTCTTTTTTGGAAAATTTAGTTTTTTCATTTTA
TTTTGAATGTAAATTGTATTCAAGATTTGATTTTGTTGGTGGGTTTGGAG
ACCCTTTTGGATTTTAGTTTCAGTTTTGTATTGTATTGGAAATGGGTGGT
GGTTAAAAAAGAGAAAATTGAGTTTGGGTTTTGTGTTTTGGTGGTGCAGT
GGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCATTAGC
TCTGTGTCCGGCAACTAAGAACTGCGTGTCAACCTCTGAGAATATCAGTG
ATCGCACACATTATGCTCCTCCATGGTTAAAGTTCCCCTCTTTTTCTTAT
TTTAATTTTCACCTTCGATTTATGGGATTATATGAATTAAATGCAATTTT
TTTTACTTGTCTGTTTGGATAAACAACTTAGTTAAGTATTCATCATGTAA
GTGCTTATGTATAAGTTGTTTCTATAATAAATAAAAATGAGGAGGGAAAG
TTATTCCAAAAATCACTTTAAAAGAGGTACTCACTTTATTTTAATTATTG
ATTTTTTTAAAATTTAATGATTAAGATTAATTATTTATTATAATTATTAG
ATTCTAAAAAAATAAATAATAAGGATAATAAATAACTCTAAAAAAATTAT
TCTTAGAGCAAGTTGATGTTTTCCTAAAAAAAATATATAAATTGTTTATA
TAAGTCGTAAGCTTTTCGGAAATTATTCTTGAAATCTTATTGAAATAATC
TGAAAACAACTTTTTTTTACATGATTTGAAAACAACTTATAGACATATCA
TAATCACATATCATTAAGTTATTTTATTAAGTCATTTTCATAATTTATTT
CAAACACTTACATAAATACTTATAAGAGAAAATAAAATAAAATTAATTTC
TTTATAAGCTATAAAATTAGTTAATGTATAAGTCAATAGGTAGAAGCTCT
CTCGTATTAACTCTTCAAAAGTTTATTTTTTAACTTATATATAAGCTAAA
TTTAACTTAAAAGAGAAACTTATTTCATTTTTTCTCTTCCTTTCTTTTCT
AGTAAATGTTTTTATAAAAGTTTACCCAAACAGAACTTTACACAAGTACT
TATAAGATAAGTCTAATTAAGCTTTTTCAAACATGCTCAAAGTTAAAGTG
TTTCCATAATGTTTTTTTGGGCCTTAATTTTCGTCAAGTAACTTGTGATG
TCAAAATTGCGTGTTGAGGAGGGTCTGGTCAACTAGATTTAGGAGATAAT
TAAGCAGTCAGGGTTTATTTATTTATTGGATTTAATCCAAAAGGAACCTA
TGACATATTTGTGAGAATCACGTTGCAATAGATAATGGTGCACTTGGAAC
AATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAACCGATATTGGACT
ATTGACTGTTGAACATTGTTGACTTGACATAAGAAATGAACTATTGGTCA
CACACGCATTGGCCGGTGAAAGTAGTGCAATCTTTACTTTTTCTTTTTTT
GAAATTTTTTTTTTCACTACCTTTGGTCCTTGTATTGAGCATGGTCCCAC
CAAAATCCAAACTTTATTATTGGACATAGATGAATCATGATGTCACTTTG
TTTTAATATTTTCATTCTTTTTCAGGAACTATAATCCTGAAGGTAGGAAA

AAACCTGTGAGCAGGGAAGAAGCAATGGAGGAACTTATAGACGTGGTAAT
AAATGCAACTGAACTGAAATCTTGAGTTATCACTGGATTGAAATTTTCTT
TTCCTTCCCTCATTTTATCAACATTGATTATAATTTATAAAATTTATGAA
AGGAGTGATAGTGAATATACACTTTGTAACACTATTTCTAATACACTTTC
TATTATCGGTTAAAATTTATTGAAAACAGTGTTGTTAAATGGCGGCCATG
GCGGCGCCATGGCGGAGTTGCGTAACGGTTTTCTGAAAAAACGCCACCGA
ATAACGGTGGCGTGGCGGATTAAAGATGGCGGCGCCATGGCGGCCGCCAT
AGCCATGGCGGCCATGGCGGATGTGGCGGGGAGGCGGAAAATGGCAGAAT
TTTTTTTTTTGTCCGCGGTAGGAGTTGGGCTGACCCGATCCAACCCTACC
CGAAAACTTAATGAAAACCATGACCCCCCCTACCTTTCAGAACGCTGCTG
CAACCCTCGAAGCTTCAACATAGCGCGACCTCGGAACCAGCCACGACCCC
TGCAACCAGCCTCGGAACCAGCCGCGCTTGACCCGCTGCACACAGCAGCC
AGCAGCGACGACCCATGGCGTGAACGGCGGCGACGAGCACGGCGTGAACA
ACGGCGACGCGAACGGAGAAGACGACCCAGAACCGCGACCTCGACTTATA
CGGGTGGGTGGGCCAAAGCCTTTTTTTTGTTTTTGCTGCTTGACACCCCT
TTTTTATGTCTGCAGCTTTTTTTCCGTTTTTCTATTTGACACCCCTTTTT
ACTTTCGACAGTCCCATTTTTAATTTTTTTTTCTATTTGACACCACAATT
TTTTTTTCTGTTCAGTCCTCTTTTAAATGGCTGAACCATCATCCTCTTTA
ATGAGTTTATTTGGTGTTGGTACTTGATTATTGTATGAACTCATGAAACT
TTTTTAGTTTATTTGAATGCAATCCTTTGTTTTTTTCAATTTCAATGAGT
TTATATATATGTTTTTTTTTTTTTTTGGTCCGCCATGACTTCCGCCATTT
TCCGCTACGCCATCCGCCATATTTTTATGGCGGATTTTTTACTTTCCGCC
ATGAACCGCCATCCGCCATTAACAACATTGATTGAAAATTATGAAGTCAT
GAGAGGAGCTCATTGAATAAAGAGTGAGAACTTACATGATTTTGTAATTT
CTAATAATTTTTTACTATTAATAAAGAGTGTATTTAAATGGGTATGTTTT
TGAACATTTTTCTAATTTCTAATCGATTTGTAATAATAGCTGGTTATTTG
CACTTTCCCAATCATTGAAGTAAAGTTAAACCAGTTCCGGATAATTTTAC
AGATAGAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAGG
AAAGAAGACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGGTAAG
TGTAACTTACATCTAAGGAAACTCATCATGAAGAAAAATTATCCTTTTAT
ACATTTTAGATGATATCAAGATTCAAGAACCATCTTTAATTTCCTTCTCC
CTTTTTTCTGTCATGTGCTAACAGTTTGTGGATGATGTTGAGTTCTGGTT
TCCTCCGGGTAAGGGTTCTACTGTGGAGTATCGTTCTGCATCTCGGTTGG
GAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATGATTCCAT
AATTCATATGTGCTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCAG
GGTAAAACTCCCCTTTCCTTCATGTCATGTGAAGCATTTTTTACTCAAGT
AGATCCACTAAATTTGAGTCTCAAATGTTTTAACTTTATTCTAAATGTTT
TAACTTTATTTGAGTCTCAAATGTTTTAACTGAAGGTAAATTTGGTTAAC
TATGATCAGAAATACATTAACAAGAGTTGAAGAATATTTTTTATGTACAA
TAAAGAGAGTATTTGCTCGAGAGAATGTAAATCCTTTTCTAAATATTTTT
GTGATGAAAAATAATGGTTGCTGGCAGGCACTGAGACAAGAGTTGGAGAA
GAAAGGATGGGCATCTCAAGACACCATATGATGAAAAAACTTAGGCAGAA
TTCACATCAGCATCTAAGAAAATATTGTTTCATATACATTGTAACCTTGT
ATACTTTTGTATTAGATACAAAATCTCACAAGATCATTGAAAGCAAACTC
TTCATGATTATTGGAATTGTAGAAATGATTGAGAACAGTACTTCAAACTC
TCGGGGGAAGGAATGAAATGAAGATGTTACCCATATCTTTTTGAACTTCA

LJFgene8

TTATCTTTCATCTTTTTAGTATCTTTACAATCTATTTTTATATTAATATT
TGTAATTTATTGCAAAAAATTTAGTAATTATAATATGGATTTTCTATGAA
TATACTATTGCATACTATAACATTAGTATTATTTACTAATTACAAATAAA
ATGCATATAAATAAAGAGTTGGAAATATTATTTCTAAAACAGCATGATAT
ACCTCTAGTGTCCACTTAATGGTGAGATTTGAAATAATATGTATGAGATT
CTAAGTTCAAATATTAAGTGTCATTGTTAAAGAAAAAACAGTATGATATG
CTGGAATTTACTAACTATAATCTTTATAATAAGTTTAGCATTTAGTGTAT
ATTGACTTGAAAGATTCTTGTAGCAATTTGCAGCAGTTTATATAGATAG
TAACATTCTTAAATTACGTTTATAAGTTCCTTCATTTTCAGCTCCGAACA
ATGGCTTCTTCGTTCTCCTTCTGCACCCTCAAGTTTCGCACCAAACCCAA
CGATAGTAGAAGCAGTGCTTCCTCTCTTCCCCGTATTCTATTCTGTCACA
ACCTCCACGATGACATTCACACACCCACTGACCAAATCAACCGAAGGTTC
ACTTCTCCACACTCTCATGCACTTTCTCACAGCTTTCTAATTTCTATTGA
TTATTCATAATTATTCATGCATACCCATCTTCTGAAATCTCTTTACACGT
CAATGATTTTGTATCTTAAAGACAACTCATATTGAGAAGCAGCGAAATAG
CGACCATCGGTGCCATCTTCGACTTCAGGTACCCCCGGCCCCTCCTCTGT
TTTTTAGAAATTTTTCTTTTCATTTTATTTTGAACGTAAATTTAATTCAA
GATTTGATTTTGTTAGTGAGTAGTGAGACCCTTTTGGTTTTTAGTATTAG
TTTTGTTTTGTATTGAAAATGGGTAGTTATTAAAAAAGAGGGTATTGTGT
GTTTTTTTTTTTTTTTTTTTTTGTGGTGCAGTGGGAAAAAACCTGATTAT
CTTGGAGTGCAGAAAAACCCACCAGCTTTAGCTCTGTGTCCGGTAACTAG
GAACTGCGTATCAACCTCTGAGAATATCAGTGATCGCACTCATTATGCTC
CTCTTTGGTAAAAACTTTCCTTGTTTTCCTCATTTTAATTTTAGCCTTCC
TTTTCAAACCGCCAAGTTAATTTTTTAACCGTGTATTGGTGGTTTTTTTT
TGGTCCTTGAATGATACATAGGATTTTTCAAATAATTGGTTCAAAGACTA
ATTATATTTATAATATGTGATTGTAGTTGGTCAAAGGGGAAATTGTTTGA
TTTTTATCATGTATAAAATGTTTTTGGACCAATAATAATAACAAATAAAA
TAAGATTAATTTTTAACTCAATTGGTTGATTAAGAGTAAAAATATTTGTG
ATCTCTCATCAAAATAAAAAAATTTCTCTTCATTTTAAAACAAAAATCAC
ATCTTTAAAATTTAAAAGTCAGGTCCACACTTAAGATACTTAACTTAAAC
ATAGGTACTCAAACCAAAAATTTTCAAATATGTGAGATCCCACACAAATT
TTTATTTATTAGAAATCTTAAAATAAATTAAATGTTTATCAGGAATTTGA
ATGTATTTTAGCCTATTTTCGAACAAGGTTTTATCATACAAAATTTATGG
ATCCAACAAAATAAAAATAAGAAATTGTCATGGCTTGGTTTCCAAAACTG
AACCTCATGTTCAAGAACACGTGGAAGCACAACTGCAGCTTTGTCTTAGC
TCATCATCAAGCGCCAGTTAGCAATATTTTGCGGGAATAGACAATGAGAC
CAAAACCTTCATCATGCACAAGTCTATTTTAGTTGCACTAGGGTTTTGGA
GCCTTTCACAAGACCAAAACTAGATACGATTCATGCTAAAATAAAGCCTA
AAGCCTTTTGGAATTATTCACAGCAAAATCCGATGTTGACATCAACCATA
TGTAACATTATAATATATTACTAGCCTTTAAAATTCAATTCCACATTCAA
CAAAATTTATTACCCTGTTCTCTCTACACTACGCAATTAGAAAACAAAAG
GTTGAGCGGAGAAAAATCTAGAAAGTGCTTAGAAATTCAACTAAATTTTG
TCAAACAGGACGAGCCAATAAATACGATTAATGACAGCTACGAGAAATGT
ATATTTTTAGTTTAGAAACACGTCTTTAGTTATATCATAACAAAAAATAA
AAATTAAAGTTGTGTTTAAATATATGTTTTATTTAAAAAAATTAAAAATT
AGGAAATTGTGTTTTAATTCATGACTTTAAGTTAAAGTCACGAATTAAAA
GATGGCTTTAATAAAAGTAAATTTAAAAAATAAATAAAGAAGTCATGTTT
GAAAATATACCTTTCGTTGAAATGTTGTGCTTTTAAAACACAACTTGCTC

ATTTATTGATTTTTTAAAAATAAATTTACCTATTTTGGGATTTTTTTTTT
ATAAAAAATGCACCACTCTAGTAAGTAAACCTTTTTTTTATATATAAAAA
ATTGCATTCAGGTAAACAACATAATTAAGTGTTTACTGATTCATTGAAAC
ACTTATGTATAAGTTGTTTATGTGATTGAAGAGAAAATAAAGTTAAATTA
TTTTCTTATAAATTGTAATATGTTTTCATGAGCTATGGAAAGTTTATTGA
AATAAACTGAAAATAGATTGTGGATATTTCATAAATACATATCTTAAAAT
TATTTTAAATATTCCAAACACTTATATAAATACTATAAGCACTTGTAATA
GAAGAAAGATAAGAATGTAAAATAAATTGATTTTTTTTCCATAAGTTGAT
TTAAGTTAAAATCAACTTATGTATCATAACACCTCAGGTTTTGAAAAAGT
TAAATGAGAGAGTTTTTAACAAAGTTAAGTGTATAAGTTAAATTTAAGAG
AAAAACACAATTTATTCTACCTTTTCTTTTCCCCATTTGTAATTGTTTAT
GGCCAATTTGATCCAAACAACATCTTTAGCCTAAATAAGCTCTTCCAATC
ACACTCTAAGTTTAAGTATGATTACTATGATGTTTAGATTTCATTGTGTT
GTTTTTTATGACCTTAATTTTCCGTCGAGTAACGCCATATTTGAATTAAA
GTTTGTAAATTTTAGATGAAATTCATTTTGCAAACTGAATTTGTAAAATC
AACTCTTCCTTTACTTCCAAAATCCAAACTTTATTATTGGACAAAGATGA
ATCGTGGAGTTACTTTGTTTTAAATTTTTAATTTTAATTATTGTTCAGGA
ACTACAATCCTGAAGGTAGGAAAAACCCTGTGAGCAGAGAAGAGGCAATG
GAGGAACTGATAGACGTGGTAATAAATCTAGCTGAAATCATAAGTTATTT
TCATGAATGCATTGCAATTTCCCTTTCCTAGCCTGTGTATCAACAAATGT
TATTATTTATAATAAGTTTAATTTTCATGCACTGACCGTATATAATAATT
TTATATTGATATCTAATCATAAATCATCATTTAAATTATTTTAAGATAAT
TAATTTAAAAGTTAATAAATTTACCGTATATAATGAATTATAATTGAATA
ATTGTATAAAAAATTTATAATGTCACTGTATAATTCTTTTTCTCATTTAT
AATTTGGTTGTAAGTTGTAGCTTGTAATAGCAGGTTATTTACCTTTTTCC
AATCATTGAAGTAAAGTTAAAATCCAGCTCTGGATAATTTTATAGATAGA
ATCAACAACACCAGACAAATTTACACCACGAATAGTTGAAAGGAAGGAAG
ACTATATTCATGTGGAGTACCAAAGCTCAATCTTGGGGGTATGTGTAACT
TACATCAAAAGGAAACTCATCGTGGAGAAAAATAATAATTTTGTACATTT
TAGATGATAATCAAGAACCATCTCTAATCCCTTCTCCCCTCCTTTTTATT
TTTTCTGCCATGTGCTAGCAGTTTGTGCATGATGTTGAGTTCTGGTTTCC
ACTGGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTCGGTTGGGGA
ACTTTGATTTTGATGTGAATAAGAAAAGAATAAAGGTATGTTTGTATCAT
TCCTTTGTGCTGTCTCGGTAGTTAACATGAAGAAATGATTAAAAGATATT
TTGTCCTTTAGGTTTTTTGGTTATATTTAGTTTGATTTTTTATTTTTTAA
AAGTTAAATTTAGTCCTTTATGTTTTTAAAACGAATCAAAATGATCACTA
TTTCGAGATGTAAAACTTTTTTATTAGATTTACTTACATGCAAATTATGC
AACTTGACTAATGTTTGAAAAAAAAAATCATCAATATGATAAAAATAATG

LJFgene9

CACTTTGTAGCACGTTTTCCAACTTCAACATTCACATATTAAAAAACAAC
AAGGGTTCCTTTTCTCGTCGATTTCAACTCTCTCAGAAGCTGGATGACGA
TAATTTCATTGATAAAATCAAACGAAGGTTCTCACTGATTCTCCCTTTAA
TTTGCCACCTCACATGAATTGTATAATATATATTTATATTTATGCTTGAC
CTTGAATTGTTCCTATCTTAAAGAGAGCTCATACTGGAAAGTGGAGAATT
AGCAACCATTGGTGCCATCTTCAACTTTAGGTACACTGCTTTATTGTTTT
CACAATGAGAATAGGTGACCAAAATTTAATTAATTCTTATAATATTTGAA
AATATTTCGAGAAATTTTATGTAAAATTAACTCTTTTTCTTTAAGATTTA

TGTGATTACTTCATAGCAAATTCTTGTCAAGTTTTAAGAACTTTTGAGTT
TTGATGTGTGTTTTCTATTGGAAATATGTGGGGATATATTATTTTGGATT
TTATGCTTCTTTGCAGAGGCAAAAAGCCAGATTATCTTGGAGTGCAGAAA
AATCAACCGGCATTAGCACTATGTCCGGCAACTAAGAACTGCATATCGAC
ATCTGAAAATGTCACTAACCTCACACATTACACTCCTCCTTGGTGAAATT
TCCTTCTTTATTTTTTTATTTATTAAAGTTTTAACTTTGGTTTATGATAT
TATCTGAATCTGAATTGGCTGCAAAGCCTGCAATTTGTTATTGAATATAT
ATTGTTGTTGAGTTTAATTTTACACTGAATAACTTTTAAAATAATTATTA
TAAAAATCAATAAAATTATTATCTATATAAAAAATATGATTAGATGATAA
AGTAAAACTTTTGTATGGTATTAATACATGAATTATTTTTTGTCCAGAAT
TTTTCAGATTGTCATAAATTTGCATTTTGATAAAGTGTATAATTCACTCA
ATTTCAAAATAACTGACTTTTTAACACCAATTTGGCAAAATGTTGCTTAT
GGAACAATTTAATTAATTATCACATTTTAAAACCATTGAAAGTTCAATTG
GTGTTGCCGTGTTGGGCGCATTGAGCCCTTAACTTAGGAAATGTATTATC
TCATGTTGAATTTGCTTTTTTTTTTTTTTTTTTATTTCTCTTCAGCTAAGTT
CCGAGCATTGAGCCTTTAACTTAGGAAATGTATTATCACATGTTGGGTAT
AGATTATTAATTAAGAGAACATTGATTATTCTACAGGAACTACAATCCTG
AAGGTAGGAAAGATCATGTGAGCAAAGAGGCAATGGAGGAACTGATAGAT
GTGGTAATTTTAATTAGGATCATGTTAAGTCTTAAGCTACTTAGTTAAAG
AATCAATAAACAATTTTTTTAGGAAATCACGTGATTAATTACATTAGAAA
TCACAGTTACACTGGCAATACTGTTATATAAAAACTATTTATTTTAAATT
GGTTGACTAGTATCTTAAGAACACTTATTAAAAAAGTGTATTGAGTTTTT
GTATCTCAATAGGAGTATAATTAAGACTATTTATAGTTGGTTTGTTATAG
CTACTGATGAGTTTTCAAATCATTAAAGTTACAAAAATCAGTTACTATTG
TTTTTTATTTTACAGATAGAATCGACAATACTACCAGAAAATTTTACACC
AAGGATTGTAGAAAGAACAGAAGATTATCTTAGATTGGAATACCAAAGTG
TATACAAGCCACAAATTTTAACTTCAATGTCACCAATATCATTGTATGCA
GAAAAAATGAATAGTAACTTTTTACTATTAGACTGAAAAGCCTGCATCAA
GCATTGAAGGAATGGAGTTTCTTTGATCATTGGACTGCCCATCTCATAGT
AACTCATCTTGAAGCAATTAATGCAGTAAAAACTAACTCATCGTGAAAAG
TTCATTCTCTGCTTTATTTAAATTTTTACAGCAAGTAGATTGGAATGCAT
GGTTTTTGCCATGTTTTATACTTGACAAAGATAATGCAAACTATAAACAC

APPENDIX F.

EST LIBRARY

03g02720			
Accession #	Cultivar	Tissue	Treatment
AI938349	Williams	Leaves, 2-3 week old seedlings, greenhouse grown	Complementary DNA was synthesized from mRNA using a 3' anchored poly (dT) primer. EcoRI adapters were ligated to the blunt-ended cDNA fragments followed by digestion with EcoRI and HindIII. The cDNA fragments were directionally cloned into the EcoRI-HindIII restriction site of the pT7T3-Pac vector. The ligated cDNA fragments were transformed into DH10B host cells (Gibco BRL).
BE610616	Clark	whole seedlings of greenhouse grown plants	Complementary DNA was synthesized from mRNA using a primer consisting of a poly(dT) sequence with a XhoI restriction site and a 3' anchor. EcoRI adapters were ligated to the blunt-ended cDNA fragments followed by XhoI digestion. The cDNA fragments were directionally cloned into the EcoRI-XhoI restriction site of the pBluescript vector. The ligated cDNA fragments were transformed into DH10B host cells (GibcoBRL).
BM176973	Williams 82	seedlings induced for HR (hypersensitive response)	Vacuum infiltrating plant tissue with Pseudomonas syringae pv. glycinea carrying the avrB gene (Genetics 141:1597-1604). Plant tissue (expanded unifoliate leaves) was collected at 2, 4, 8, 12, 24, 36, and 53 hrs after inoculation and their mRNA pooled equally for cDNA construction. Complementary DNA was synthesized from mRNA using a primer consisting of a poly(dT) sequence with an XhoI restriction site. EcoRI adaptors were ligated to the blunt-ended cDNA fragments followed by XhoI digestion. The cDNA insert is protected from XhoI digestion via methylation during first strand synthesis. The cDNA fragments were directionally cloned into the EcoRI-XhoI restriction site of the pBluescript vector. The ligated cDNA fragments were transformed into E.coli ElectroMax DH10B host cells.
BM886799	Williams 82	Leaf, drought stressed, 1 month old plants, greenhouse grown	The month old greenhouse grown plants were deprived of water for 3 days prior to harvesting the stressed leaf tissue. Complementary DNA was synthesized from mRNA using a primer consisting of a poly(dT) sequence with a XhoI restriction

			site. EcoRI adapters were ligated to the blunt-ended cDNA fragments followed by XhoI digestion. The cDNA fragments were directionally cloned into the EcoRI-XhoI restriction site of the pBluescript vector. The ligated cDNA fragments were transformed into DH10B host cells (GibcoBRL).
CA800126	Raiden	stem tissue of greenhouse grown plants	Complementary DNA was synthesized from mRNA using a primer consisting of a poly(dT) sequence with a XhoI restriction site. EcoRI adapters were ligated to the blunt-ended cDNA fragments followed by XhoI digestion. The cDNA fragments were directionally cloned into the EcoRI-XhoI restriction site of the pBluescript vector. The ligated cDNA fragments were transformed into DH10B host cells (GibcoBRL).
CO983876	Not given	not given	The library Gm-r1089 is a sequence-driven, reracked set of 9,216 low redundancy clones selected from 38 different cDNA libraries constructed from various tissues and stages of development of soybean including 973 cDNAs from germinating cotyledons(source library Gm-c1069, Gm-c1076, and Gm-c1077); 1,465 cDNAs from various tissue and organ systems of the adult plant; 476 cDNAs from adult stem tissue (source library Gm-c1062); 1340 cDNAs from tissue culture derived somatic embryos (source libraries Gm-c1036 and Gm-c1075); 2918 cDNAs from hypocotyls or young seedlings; 742 cDNAs from germinating seedlings, shoot tips, or leaves exposed to various stresses(source libraries Gm-c1065, Gm-c1066, and Gm-c1068); 839 cDNAs from young leaves or hypocotyls exposed to bacterial and fungal pathogens (source libraries Gm-c1072, Gm-c1073, Gm-c1074; and Gm-c1084); and 463 from roots of young plants grown in hydroponic media without phosphate (source library Gm-c1087). The 5' ESTs of the source clones from the different libraries were used to select singletons, or a representative of each contig, which were reracked to form library Gm-r1089 and the cDNA clones of the reracked Gm-r1089 library were then sequenced at the 3' end. The unigene selection and 3' sequencing

			was funded by NSF Plant Genome project #9872565.
EV275072	Williams 82	drought stressed and control stem, tissue culture suspension, salt stressed, drought stressed, and Pseudomonas infected leaves, etiolated seedlings	Total RNA isolated using TRIzol reagent. No DNaseI digestion was performed. Total RNAs were pooled together. Double strand cDNA were synthesized from pooled RNA using SMART technology (Clontech). The prepared cDNA was normalized by cDNA denaturation/reassociation, treatment by duplex-specific nuclease(DSN) and amplification of normalized fraction by PCR. The normalized cDNA was then digested with SfiI, size-fractioned, directionally ligated into pDNR-LIB (Clontech) and electroprated into GC10 competent cells (Gene Choice). Clones were sequenced from the 5' end only.
14g07660			
EV264523	Williams 82	stem, apical meristem, flowers, young pods, green seeds	Total RNA isolated using TRIzol reagent. No DNaseI digestion was performed. Total RNAs were pooled together. Double strand cDNA were synthesized from pooled RNA using SMART technology (Clontech). The prepared cDNA was normalized by cDNA denaturation/reassociation, treatment by duplex-specific nuclease(DSN) and amplification of normalized fraction by PCR. The normalized cDNA was then digested with SfiI, size-fractioned, directionally ligated into pDNR-LIB (Clontech) and electroprated into GC10 competent cells (Gene Choice). Clones were sequenced from the 5' end only.
FG993792	Williams 82	stem, apical meristem, flowers, young pods, green seeds	Total RNA isolated using TRIzol reagent. No DNaseI digestion was performed. Total RNAs were pooled together. Double strand cDNA were synthesized from pooled RNA using SMART technology (Clontech). The prepared cDNA was normalized by cDNA denaturation/reassociation, treatment by duplex-specific nuclease(DSN) and amplification of normalized fraction by PCR. The normalized cDNA was then digested with SfiI, size-fractioned, directionally ligated into pDNR-LIB (Clontech) and electroprated into GC10 competent cells (Gene Choice). Clones were sequenced from the 5' end only.
HO044862	not listed	immature seed	not listed
09g40540			

CD399608	Kefeng 1	two-week seedlings	Spraying 2.0mM salicylic acid for 24, 36, 48 and 72 h. Complementary DNA was synthesized from mRNA using a primer consisting of a poly(dT) sequence with a XhoI striction site. EcoRI adapters were ligated to the blunt-ended cDNA fragments followed by XhoI digestion. The cDNA fragments were directionally cloned into the EcoRI-XhoI restriction site of the pBluescript vector. The ligated cDNA fragments were transformed into XL1-Blue MRF' host cells (Stratagene).
CF921901	Williams 82	root hairs	cDNA clones generated from soybean root hair tissue treated with Bradyrhizobium japonicum for 24 hours.
CF923165	Williams 82	root hairs	cDNA clones generated from soybean root hair tissue treated with Bradyrhizobium japonicum for 24 hours.

APPENDIX G.

PLACE FULL OUTPUT

Segment of chromosome 3 upstream of LJFgene3 start codon:

TTATAAAAAAGTTCTCCATCTCCCTAATTTTTGCATCTCCAAAACATTCT
CTCATTATATATAATTGTCGGTTCTCACTTTCAACAACAACTTTTTCTCA
CATGCACATTCCATTCATATTGGGTGTGAAGAATGCGTTATGAATTTCGA
ACCACAACCTCTACGTGTAGATGTAACTTTTGAAAGTTAGAGTTTAATGT
ATTTGGGGGAGATATTTTAATAACTTTTAAAAAATAAATAAATAAAATTC
CCATGTTATTTTAATTTATTTTTTTAAAAAAAAAACACACTGTTTGCCAC
GTTGAATACAAGCAAACACCATTAGAACACAAGTTAACATCGTTAGTGAA
AATTCTGTGAAAGATGATTAAAAATAAGTTTTTTTAAAGTTTAGGGACTA
AATTAGAGGATATAAAAGTTTAGAAACTAAAAACAAGATTCAATGAAAAT
ATAGGGATTAAAAAATACTTTCCCTTTTTTTAAATGACATTTTCAAAATA
TGTAAAAAGTAAATAATTGTTAAATCTAAGGACATGTTTGGGTAATGTTT
TCTAGGAATACTTTTAAGAGAAGAAAATAAGAAAAAAATAAAATAAAATG
AGTTTCTCCCTAAGTCAAAGAAAGCTCTCCATTAGTTAATTTATAAAAGT
TCTCTCATCTTTTAAAAAAGTTATATGAGAAAGTTTCTACAAATTAATCA
TTTTGATTTATGAAAAAATTCATTTTATCTTTTTCTTCCTATTTATTTTC
TCTTAAGAGTACTTCTAAAAAAAACTTATCCAAACAAACTATAAATATAC
ATATGAAGAGTTTACCATACATATAAAAAAATGGAAGGGTTTATCTATCT
TAAACTTAATATTAGAGTACATGTATTTAAGTGTTATTTTATTTAATATG
TAAAATTATATTATTAATAGTACATGTATCTTTTTATCTAATATGGTTTT
TTTATTAATATTTCAAATGAGTATTATTTAAGTTTTCAATATAATATTGA
TAATAAAATTAATTTATAAAATTATTATTTTTATCTATTTTTTATTTAAT
GTAAAATAATCTAGGACAACCCAAACAATTAATATAATGGTATAAAATTT
TATGCATTCATATGGGGAGAAATTAGTATTTGATTCATAAGGTGTGTAAA
ATCTTCTTTGACCAATGACCACAAAATACTTTTAGTCAAATAAGTGATTT
GGAGACAACCATAATTTCTAATTCAAGAAAATAAAATCTAAAATGAAACT
TATTCTAAAAAGTAAACGGATATAATAACATCCTTCACAAAAACCTTTTC
GTGCATATAAGCATTTCTGGATATTTGTATTATCTGAAATTGGAAATAGT
CCACTATCTTAAAAACATCGAAATATAATTTTTTTTCCAAATATCCAAAT
CCATAAAAAAGAAAAAAAAAAAGTCCCACCGCCCACCTTCTTTATCACAT
GATTCACATCTCATTCCTTATATTTGGTTCACATTCTTAAATTATAAATA
TTTCGGTCTGTGAAGATATATGTCCATAAGTTCCTTAATTTTCTCGAACC
TTCATTTTCAGCTCCCAACAACAATGGCTTCAATGGCATCTTCAAGCTC

Element name	Site/Strand/Sequence	Identifier link
-10PEHVPSBD	1251 (+) TATTCT	S000392
-300ELEMENT	899 (+) TGHAAARK	S000122
-300ELEMENT	1050 (+) TGHAAARK	S000122
-300ELEMENT	1145 (+) TGHAAARK	S000122
2SSEEDPROTBANAPA	313 (+) CAAACAC	S000143
AACACOREOSGLUB1	783 (+) AACAAAC	S000353
ABRELATERD1	163 (+) ACGTG	S000414
ABRELATERD1	298 (-) ACGTG	S000414
ABRERATCAL	297 (-) MACGYGB	S000507
ACGTABREMOTIFA2OSEM	296 (-) ACGTGKC	S000394
ACGTATERD1	163 (+) ACGT	S000415
ACGTATERD1	299 (+) ACGT	S000415
ACGTATERD1	163 (-) ACGT	S000415
ACGTATERD1	299 (-) ACGT	S000415
AMMORESIIUDCRNIA1	833 (+) GGWAGGGT	S000374
ANAERO1CONSENSUS	782 (+) AAACAAA	S000477
ARFAT	1202 (-) TGTCTC	S000270
ARR1AT	436 (+) NGATT	S000454
ARR1AT	455 (+) NGATT	S000454
ARR1AT	365 (+) NGATT	S000454

ARR1AT	704 (+) NGATT	S000454
ARR1AT	1131 (+) NGATT	S000454
ARR1AT	1195 (+) NGATT	S000454
ARR1AT	1450 (+) NGATT	S000454
ARR1AT	523 (-) NGATT	S000454
ARR1AT	696 (-) NGATT	S000454
ARR1AT	1058 (-) NGATT	S000454
ARR1AT	1150 (-) NGATT	S000454
ARR1AT	1235 (-) NGATT	S000454
ARR1AT	1398 (-) NGATT	S000454
BIHD1OS	485 (-) TGTCA	S000498
BOXIIPCCHS	296 (-) ACGTGGC	S000229
BP5OSWX	298 (-) CAACGTG	S000436
CAATBOX1	441 (+) CAAT	S000028
CAATBOX1	987 (+) CAAT	S000028
CAATBOX1	1076 (+) CAAT	S000028
CAATBOX1	1163 (+) CAAT	S000028
CAATBOX1	1572 (+) CAAT	S000028
CAATBOX1	1581 (+) CAAT	S000028
CAATBOX1	64 (-) CAAT	S000028
CAATBOX1	119 (-) CAAT	S000028
CAATBOX1	516 (-) CAAT	S000028
CAATBOX1	996 (-) CAAT	S000028
CAATBOX1	1339 (-) CAAT	S000028
CACTFTPPCA1	76 (+) YACT	S000449
CACTFTPPCA1	288 (+) YACT	S000449
CACTFTPPCA1	1352 (+) YACT	S000449
CACTFTPPCA1	466 (+) YACT	S000449
CACTFTPPCA1	559 (+) YACT	S000449
CACTFTPPCA1	760 (+) YACT	S000449
CACTFTPPCA1	1177 (+) YACT	S000449
CACTFTPPCA1	345 (-) YACT	S000449
CACTFTPPCA1	508 (-) YACT	S000449
CACTFTPPCA1	758 (-) YACT	S000449
CACTFTPPCA1	866 (-) YACT	S000449
CACTFTPPCA1	880 (-) YACT	S000449
CACTFTPPCA1	919 (-) YACT	S000449
CACTFTPPCA1	970 (-) YACT	S000449
CACTFTPPCA1	1125 (-) YACT	S000449
CACTFTPPCA1	1193 (-) YACT	S000449
CACTFTPPCA1	1261 (-) YACT	S000449
CANBNNAPA	313 (+) CNAACAC	S000148
CARGCW8GAT	24 (+) CWWWWWWWWG	S000431
CARGCW8GAT	1402 (+) CWWWWWWWWG	S000431
CARGCW8GAT	1467 (+) CWWWWWWWWG	S000431
CARGCW8GAT	24 (-) CWWWWWWWWG	S000431
CARGCW8GAT	1402 (-) CWWWWWWWWG	S000431
CARGCW8GAT	1467 (-) CWWWWWWWWG	S000431
CARGNCAT	1466 (+) CCWWWWWWWWGG	S000446
CARGNCAT	1466 (-) CCWWWWWWWWGG	S000446
CATATGGMSAUR	800 (+) CATATG	S000370
CATATGGMSAUR	1109 (+) CATATG	S000370
CATATGGMSAUR	800 (-) CATATG	S000370
CATATGGMSAUR	1109 (-) CATATG	S000370
CBFHV	67 (-) RYCGAC	S000497
CCAATBOX1	1162 (+) CCAAT	S000030

CCAATBOX1	119	(-) CCAAT	S000030
CCAATBOX1	1339	(-) CCAAT	S000030
CELLCYCLESC	1296	(-) CACGAAAA	S000031
CEREGLUBOX2PSLEGA	981	(-) TGAAAACT	S000033
CIACADIANLELHC	690	(+) CAANNNNATC	S000252
CIACADIANLELHC	1581	(+) CAANNNNATC	S000252
CPBCSPOR	860	(+) TATTAG	S000491
CPBCSPOR	938	(-) TATTAG	S000491
CURECORECR	759	(+) GTAC	S000493
CURECORECR	867	(+) GTAC	S000493
CURECORECR	920	(+) GTAC	S000493
CURECORECR	759	(-) GTAC	S000493
CURECORECR	867	(-) GTAC	S000493
CURECORECR	920	(-) GTAC	S000493
DOFCOREZM	8	(+) AAAG	S000265
DOFCOREZM	183	(+) AAAG	S000265
DOFCOREZM	360	(+) AAAG	S000265
DOFCOREZM	386	(+) AAAG	S000265
DOFCOREZM	415	(+) AAAG	S000265
DOFCOREZM	506	(+) AAAG	S000265
DOFCOREZM	617	(+) AAAG	S000265
DOFCOREZM	621	(+) AAAG	S000265
DOFCOREZM	646	(+) AAAG	S000265
DOFCOREZM	667	(+) AAAG	S000265
DOFCOREZM	680	(+) AAAG	S000265
DOFCOREZM	1259	(+) AAAG	S000265
DOFCOREZM	1408	(+) AAAG	S000265
DOFCOREZM	1420	(+) AAAG	S000265
DOFCOREZM	78	(-) AAAG	S000265
DOFCOREZM	91	(-) AAAG	S000265
DOFCOREZM	177	(-) AAAG	S000265
DOFCOREZM	224	(-) AAAG	S000265
DOFCOREZM	468	(-) AAAG	S000265
DOFCOREZM	474	(-) AAAG	S000265
DOFCOREZM	561	(-) AAAG	S000265
DOFCOREZM	659	(-) AAAG	S000265
DOFCOREZM	729	(-) AAAG	S000265
DOFCOREZM	930	(-) AAAG	S000265
DOFCOREZM	1156	(-) AAAG	S000265
DOFCOREZM	1179	(-) AAAG	S000265
DOFCOREZM	1295	(-) AAAG	S000265
DOFCOREZM	1440	(-) AAAG	S000265
DPBFCOREDCDC3	327	(+) ACACNNG	S000292
DRE2COREZMRAB17	67	(-) ACCGAC	S000402
DRECRTCOREAT	67	(-) RCCGAC	S000418
EBOXBNNAPA	99	(+) CANNTG	S000144
EBOXBNNAPA	800	(+) CANNTG	S000144
EBOXBNNAPA	964	(+) CANNTG	S000144
EBOXBNNAPA	1109	(+) CANNTG	S000144
EBOXBNNAPA	1446	(+) CANNTG	S000144
EBOXBNNAPA	99	(-) CANNTG	S000144
EBOXBNNAPA	800	(-) CANNTG	S000144
EBOXBNNAPA	964	(-) CANNTG	S000144
EBOXBNNAPA	1109	(-) CANNTG	S000144
EBOXBNNAPA	1446	(-) CANNTG	S000144
EECCRCAH1	142	(+) GANTTNC	S000494

EECCRCAH1	600 (+) GANTTNC	S000494
ELRECOREPCRP1	1158 (+) TTGACC	S000142
ERELEE4	960 (+) AWTTCAAA	S000037
EVENINGAT	210 (-) AAAATATCT	S000385
GATABOX	211 (+) GATA	S000039
GATABOX	409 (+) GATA	S000039
GATABOX	999 (+) GATA	S000039
GATABOX	1269 (+) GATA	S000039
GATABOX	1320 (+) GATA	S000039
GATABOX	1515 (+) GATA	S000039
GATABOX	726 (-) GATA	S000039
GATABOX	777 (-) GATA	S000039
GATABOX	842 (-) GATA	S000039
GATABOX	846 (-) GATA	S000039
GATABOX	927 (-) GATA	S000039
GATABOX	935 (-) GATA	S000039
GATABOX	1032 (-) GATA	S000039
GATABOX	1331 (-) GATA	S000039
GATABOX	1355 (-) GATA	S000039
GATABOX	1392 (-) GATA	S000039
GATABOX	1443 (-) GATA	S000039
GT1CONSENSUS	348 (+) GRWAAW	S000198
GT1CONSENSUS	445 (+) GRWAAW	S000198
GT1CONSENSUS	541 (+) GRWAAW	S000198
GT1CONSENSUS	573 (+) GRWAAW	S000198
GT1CONSENSUS	581 (+) GRWAAW	S000198
GT1CONSENSUS	712 (+) GRWAAW	S000198
GT1CONSENSUS	999 (+) GRWAAW	S000198
GT1CONSENSUS	1227 (+) GRWAAW	S000198
GT1CONSENSUS	1342 (+) GRWAAW	S000198
GT1CONSENSUS	1411 (+) GRWAAW	S000198
GT1CONSENSUS	489 (-) GRWAAW	S000198
GT1CONSENSUS	745 (-) GRWAAW	S000198
GT1CONSENSUS	1329 (-) GRWAAW	S000198
GT1CONSENSUS	1538 (-) GRWAAW	S000198
GT1CONSENSUS	1554 (-) GRWAAW	S000198
GT1CONSENSUS	92 (-) GRWAAW	S000198
GT1CONSENSUS	724 (-) GRWAAW	S000198
GT1CONSENSUS	730 (-) GRWAAW	S000198
GT1CONSENSUS	811 (-) GRWAAW	S000198
GT1CONSENSUS	840 (-) GRWAAW	S000198
GT1CONSENSUS	933 (-) GRWAAW	S000198
GT1CONSENSUS	1030 (-) GRWAAW	S000198
GT1CONSENSUS	1382 (-) GRWAAW	S000198
GT1CONSENSUS	1383 (-) GRWAAW	S000198
GT1CONSENSUS	1441 (-) GRWAAW	S000198
GT1GMSCAM4	581 (+) GAAAAA	S000453
GT1GMSCAM4	712 (+) GAAAAA	S000453
GT1GMSCAM4	1411 (+) GAAAAA	S000453
GT1GMSCAM4	92 (-) GAAAAA	S000453
GT1GMSCAM4	730 (-) GAAAAA	S000453
GT1GMSCAM4	1382 (-) GAAAAA	S000453
GTGANTG10	126 (+) GTGA	S000378
GTGANTG10	346 (+) GTGA	S000378
GTGANTG10	357 (+) GTGA	S000378
GTGANTG10	1194 (+) GTGA	S000378

GTGANTG10	1510 (+) GTGA	S000378
GTGANTG10	75 (-) GTGA	S000378
GTGANTG10	98 (-) GTGA	S000378
GTGANTG10	1285 (-) GTGA	S000378
GTGANTG10	1445 (-) GTGA	S000378
GTGANTG10	1454 (-) GTGA	S000378
GTGANTG10	1479 (-) GTGA	S000378
IBOX	775 (-) GATAAG	S000124
IBOXCORE	999 (+) GATAA	S000199
IBOXCORE	725 (-) GATAA	S000199
IBOXCORE	776 (-) GATAA	S000199
IBOXCORE	841 (-) GATAA	S000199
IBOXCORE	934 (-) GATAA	S000199
IBOXCORE	1031 (-) GATAA	S000199
IBOXCORE	1330 (-) GATAA	S000199
IBOXCORE	1442 (-) GATAA	S000199
INRNTPSADB	74 (+) YTCANTYY	S000395
INRNTPSADB	1460 (+) YTCANTYY	S000395
INRNTPSADB	719 (+) YTCANTYY	S000395
INRNTPSADB	1551 (+) YTCANTYY	S000395
INRNTPSADB	595 (-) YTCANTYY	S000395
INRNTPSADB	1240 (-) YTCANTYY	S000395
LECPLEACS2	212 (-) TAAAATAT	S000465
LRENPCABE	295 (-) ACGTGGCA	S000231
LTRE1HVBLT49	1501 (-) CCGAAA	S000250
LTREATLTI78	66 (-) ACCGACA	S000157
LTRECOREATCOR15	67 (-) CCGAC	S000153
MARABOX1	233 (+) AATAAAYAAA	S000063
MARABOX1	237 (+) AATAAAYAAA	S000063
MARTBOX	266 (+) TTWTWTTWTT	S000067
MARTBOX	884 (+) TTWTWTTWTT	S000067
MARTBOX	906 (+) TTWTWTTWTT	S000067
MARTBOX	947 (+) TTWTWTTWTT	S000067
MARTBOX	276 (-) TTWTWTTWTT	S000067
MARTBOX	582 (-) TTWTWTTWTT	S000067
MARTBOX	584 (-) TTWTWTTWTT	S000067
MARTBOX	587 (-) TTWTWTTWTT	S000067
MARTBOX	1412 (-) TTWTWTTWTT	S000067
MARTBOX	1413 (-) TTWTWTTWTT	S000067
MYB1AT	944 (-) WAACCA	S000408
MYBCOREATCYCB1	1265 (+) AACGG	S000502
MYBST1	408 (+) GGATA	S000180
MYBST1	1268 (+) GGATA	S000180
MYBST1	1319 (+) GGATA	S000180
MYBST1	777 (-) GGATA	S000180
MYBST1	1392 (-) GGATA	S000180
MYCATERD1	99 (-) CATGTG	S000413
MYCATERD1	1446 (-) CATGTG	S000413
MYCATRD22	99 (+) CACATG	S000174
MYCATRD22	1446 (+) CACATG	S000174
MYCCONSENSUSAT	99 (+) CANNTG	S000407
MYCCONSENSUSAT	800 (+) CANNTG	S000407
MYCCONSENSUSAT	964 (+) CANNTG	S000407
MYCCONSENSUSAT	1109 (+) CANNTG	S000407
MYCCONSENSUSAT	1446 (+) CANNTG	S000407
MYCCONSENSUSAT	99 (-) CANNTG	S000407

MYCCONSENSUSAT	800 (-) CANNTG	S000407
MYCCONSENSUSAT	964 (-) CANNTG	S000407
MYCCONSENSUSAT	1109 (-) CANNTG	S000407
MYCCONSENSUSAT	1446 (-) CANNTG	S000407
NODCON1GM	360 (+) AAAGAT	S000461
NODCON1GM	657 (-) AAAGAT	S000461
NODCON1GM	727 (-) AAAGAT	S000461
NODCON1GM	928 (-) AAAGAT	S000461
NODCON2GM	750 (+) CTCTT	S000462
NODCON2GM	566 (-) CTCTT	S000462
NODCON2GM	755 (-) CTCTT	S000462
NODCON2GM	806 (-) CTCTT	S000462
NTBBF1ARROLB	385 (-) ACTTTA	S000273
OSE1ROOTNODULE	360 (+) AAAGAT	S000467
OSE1ROOTNODULE	657 (-) AAAGAT	S000467
OSE1ROOTNODULE	727 (-) AAAGAT	S000467
OSE1ROOTNODULE	928 (-) AAAGAT	S000467
OSE2ROOTNODULE	750 (+) CTCTT	S000468
OSE2ROOTNODULE	566 (-) CTCTT	S000468
OSE2ROOTNODULE	755 (-) CTCTT	S000468
OSE2ROOTNODULE	806 (-) CTCTT	S000468
POLASIG1	233 (+) AATAAA	S000080
POLASIG1	237 (+) AATAAA	S000080
POLASIG1	241 (+) AATAAA	S000080
POLASIG1	587 (+) AATAAA	S000080
POLASIG1	592 (+) AATAAA	S000080
POLASIG1	1002 (+) AATAAA	S000080
POLASIG1	1230 (+) AATAAA	S000080
POLASIG1	265 (-) AATAAA	S000080
POLASIG1	742 (-) AATAAA	S000080
POLASIG1	888 (-) AATAAA	S000080
POLASIG1	951 (-) AATAAA	S000080
POLASIG1	1041 (-) AATAAA	S000080
POLASIG2	260 (-) AATTAAA	S000081
POLASIG3	512 (+) AATAAT	S000088
POLASIG3	1055 (+) AATAAT	S000088
POLASIG3	910 (-) AATAAT	S000088
POLASIG3	973 (-) AATAAT	S000088
POLASIG3	1021 (-) AATAAT	S000088
POLASIG3	1024 (-) AATAAT	S000088
POLLEN1LELAT52	422 (+) AGAAA	S000245
POLLEN1LELAT52	572 (+) AGAAA	S000245
POLLEN1LELAT52	580 (+) AGAAA	S000245
POLLEN1LELAT52	619 (+) AGAAA	S000245
POLLEN1LELAT52	678 (+) AGAAA	S000245
POLLEN1LELAT52	1118 (+) AGAAA	S000245
POLLEN1LELAT52	1226 (+) AGAAA	S000245
POLLEN1LELAT52	1410 (+) AGAAA	S000245
POLLEN1LELAT52	94 (-) AGAAA	S000245
POLLEN1LELAT52	549 (-) AGAAA	S000245
POLLEN1LELAT52	603 (-) AGAAA	S000245
POLLEN1LELAT52	684 (-) AGAAA	S000245
POLLEN1LELAT52	732 (-) AGAAA	S000245
POLLEN1LELAT52	747 (-) AGAAA	S000245
POLLEN1LELAT52	1215 (-) AGAAA	S000245
POLLEN1LELAT52	1314 (-) AGAAA	S000245

POLLEN1LELAT52	1540 (-) AGAAA	S000245
PREATPRODH	598 (-) ACTCAT	S000450
PREATPRODH	967 (-) ACTCAT	S000450
PROXBBNNAPA	313 (+) CAAACACC	S000263
PYRIMIDINEBOXHVEPB1	1381 (+) TTTTTTCC	S000298
PYRIMIDINEBOXOSRAMY1A	473 (+) CCTTTT	S000259
PYRIMIDINEBOXOSRAMY1A	1294 (+) CCTTTT	S000259
RAV1AAT	82 (+) CAACA	S000314
RAV1AAT	85 (+) CAACA	S000314
RAV1AAT	1566 (+) CAACA	S000314
RAV1AAT	1569 (+) CAACA	S000314
REALPHALGLHCB21	1474 (-) AACCAA	S000362
REBETALGLHCB21	1267 (+) CGGATA	S000363
RHERPATEXPA7	1299 (-) KCACGW	S000512
ROOTMOTIFTAPOX1	117 (+) ATATT	S000098
ROOTMOTIFTAPOX1	212 (+) ATATT	S000098
ROOTMOTIFTAPOX1	859 (+) ATATT	S000098
ROOTMOTIFTAPOX1	908 (+) ATATT	S000098
ROOTMOTIFTAPOX1	958 (+) ATATT	S000098
ROOTMOTIFTAPOX1	994 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1321 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1470 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1498 (+) ATATT	S000098
ROOTMOTIFTAPOX1	448 (-) ATATT	S000098
ROOTMOTIFTAPOX1	497 (-) ATATT	S000098
ROOTMOTIFTAPOX1	794 (-) ATATT	S000098
ROOTMOTIFTAPOX1	858 (-) ATATT	S000098
ROOTMOTIFTAPOX1	895 (-) ATATT	S000098
ROOTMOTIFTAPOX1	940 (-) ATATT	S000098
ROOTMOTIFTAPOX1	957 (-) ATATT	S000098
ROOTMOTIFTAPOX1	988 (-) ATATT	S000098
ROOTMOTIFTAPOX1	993 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1081 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1372 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1390 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1497 (-) ATATT	S000098
RYREPEATBNNAPA	101 (+) CATGCA	S000264
RYREPEATLEGUMINBOX	101 (+) CATGCAY	S000100
S1FBOXSORPS1L21	1087 (+) ATGGTA	S000223
S1FBOXSORPS1L21	813 (-) ATGGTA	S000223
SEBFCONSSTPR10A	1202 (-) YTGTCWC	S000391
SEF1MOTIF	790 (-) ATATTTAWW	S000006
SEF1MOTIF	1493 (-) ATATTTAWW	S000006
SEF3MOTIFGM	1068 (+) AACCCA	S000115
SEF4MOTIFGM7S	27 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	1027 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	1288 (-) RTTTTTR	S000103
SEF4MOTIFGM7S	369 (-) RTTTTTR	S000103
SEF4MOTIFGM7S	428 (-) RTTTTTR	S000103
SEF4MOTIFGM7S	1360 (-) RTTTTTR	S000103
SORLIP1AT	296 (+) GCCAC	S000482
SP8BFIBSP8BIB	916 (-) TACTATT	S000184
SREATMSD	776 (+) TTATCC	S000470
SURE2STPAT21	1122 (-) AATACTAAT	S000185
SURECOREATSULTR11	1202 (+) GAGAC	S000499
SV40COREENHAN	148 (-) GTGGWWHG	S000123

T/GBOXATPIN2	298 (-) AACGTG	S000458
TAAAGSTKST1	385 (+) TAAAG	S000387
TAAAGSTKST1	1440 (-) TAAAG	S000387
TATABOX2	790 (+) TATAAAT	S000109
TATABOX2	1493 (+) TATAAAT	S000109
TATABOX2	639 (-) TATAAAT	S000109
TATABOX2	1012 (-) TATAAAT	S000109
TATABOX3	912 (+) TATTAAT	S000110
TATABOX3	953 (+) TATTAAT	S000110
TATABOX3	913 (-) TATTAAT	S000110
TATABOX3	954 (-) TATTAAT	S000110
TATABOX3	1078 (-) TATTAAT	S000110
TATABOX4	58 (+) TATATAA	S000111
TATABOX4	55 (-) TATATAA	S000111
TATABOX5	256 (+) TTATTT	S000203
TATABOX5	266 (+) TTATTT	S000203
TATABOX5	743 (+) TTATTT	S000203
TATABOX5	884 (+) TTATTT	S000203
TATABOX5	889 (+) TTATTT	S000203
TATABOX5	974 (+) TTATTT	S000203
TATABOX5	1025 (+) TTATTT	S000203
TATABOX5	1042 (+) TTATTT	S000203
TATABOX5	232 (-) TTATTT	S000203
TATABOX5	236 (-) TTATTT	S000203
TATABOX5	240 (-) TTATTT	S000203
TATABOX5	372 (-) TTATTT	S000203
TATABOX5	511 (-) TTATTT	S000203
TATABOX5	575 (-) TTATTT	S000203
TATABOX5	586 (-) TTATTT	S000203
TATABOX5	591 (-) TTATTT	S000203
TATABOX5	1054 (-) TTATTT	S000203
TATABOX5	1188 (-) TTATTT	S000203
TATABOX5	1229 (-) TTATTT	S000203
TATABOXOSPAL	874 (+) TATTTAA	S000400
TATABOXOSPAL	890 (+) TATTTAA	S000400
TATABOXOSPAL	975 (+) TATTTAA	S000400
TATABOXOSPAL	1043 (+) TATTTAA	S000400
TATCCAOSAMY	777 (+) TATCCA	S000403
TATCCAOSAMY	1392 (+) TATCCA	S000403
TATCCAOSAMY	1318 (-) TATCCA	S000403
WBBOXPCWRKY1	1157 (+) TTTGACY	S000310
WBBOXPCWRKY1	613 (-) TTTGACY	S000310
WBBOXPCWRKY1	1184 (-) TTTGACY	S000310
WBOXATNPR1	1158 (+) TTGAC	S000390
WBOXATNPR1	614 (-) TTGAC	S000390
WBOXATNPR1	1185 (-) TTGAC	S000390
WBOXHVISO1	613 (-) TGACT	S000442
WBOXHVISO1	1184 (-) TGACT	S000442
WBOXNTERF3	1159 (+) TGACY	S000457
WBOXNTERF3	1166 (+) TGACY	S000457
WBOXNTERF3	613 (-) TGACY	S000457
WBOXNTERF3	1184 (-) TGACY	S000457
WRKY71OS	485 (+) TGAC	S000447
WRKY71OS	1159 (+) TGAC	S000447
WRKY71OS	1166 (+) TGAC	S000447
WRKY71OS	614 (-) TGAC	S000447

WRKY71OS	1185 (-) TGAC	S000447

Segment of chromosome 14 upstream of LJFgene14 start codon:

TAGCTAGTTAAATAATATGCTAAATTATTGCAAGTTCTTTTTGATTCCAA
CATATAGAAAATAAATAAATGCATGTATGTATCTTGCATTGACAAATTAT
AGGTATAAACCCTCACTTATTTCTTGACGTAGTAGTTTTCTTCCCTTCGG
TTATGACTCATATCCTCTCCCTAATTTAAGAATGACATAAATCTATTGTT
TATTTGAGTAGAAATGAGACCCAAGAAAAAAAAATTATGAATTGTTTGGA
AGGGGAAAAACACGAAGGAGAAAAGTTATATTTGTGATATCCATTAGAAA
AAAATTTACCTTCTAATTGAAATAAAATGAAAGGAAAATAATTTCTTTCT
ATCTTGTTTGTTTTTTATTTTTCCTTTTTTTAATTTTTATTTTCTTTCAC
AATATGTTTTTTTTAAAAAAAACGACAATTAAGTTTATGTTACATAAAAA
AATGAATTAAAAGAAAAAAAATGAAACTCAGATATATATATATATATATA
TATATATACACACACACACGAATAAACAAATTTTTTTAAGAGTAAATTAC
ATAACATCTTGTGAGATTTTAAATTTTTTTATACATATTTAAAAAAAAAA
GACTTACACAAATCTATCAATTAAATTTTAAAAAATTACACACGTCTCAT
AACTGTTTTTGAATAAATACTAACTAAAATTAAAAAAAATGTAGAAATGC
ATTATTATTTTTACCGAGTAAAAACATTCTTGATGCGCGAATTTGACAAA
AACCTTTTCGTACAGATAAGCATTTATGGATATTTTTAGTATCCAAAATT
GTCACTTTCTCAAACAATCGAAATATATACTATTTATTTTCTAAATATCT
AAATCCATAAAAAGGAAAATAAATAAAAAAATAAAAATGTTGCGGAAACG
AAGTCCCACCTTCTTTTATTCATCACATGATTCACATCTCATTTCTTATT
TTCGGGTCACTTTGTTAAATTATAAATAATTTCGTTCTGTGAAGGTACAC
ACGTTCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAAC
AATAATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCTCAAGTTTA
TCACCAAACCCAACAATGGTAGAAGAAGCTCTCTTCGCCGTATTGTATTT
TGTCAGAAGCATCACGATGACACACCCACCGACCAAATCAACCGAAGGTT
CTTACTTCTTCACACTCACACTTTCTATTTCCTTTCTATTGATTATTCGT
AACCATCTTCTGAAATCTCGTTACATTTCAATTCTTTTGTGTATTGAAGA
GAACTCATATTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTCAA
CTTCGGGTACCCCTCCTCTGTTTTTGCTCTGTTTTTTTTTCTGGAAATTT
TAGTTTTTCATTTTATTTTGAATGTAAATTAAATTCGAGATTTGATTTTG

Element name	Site/Strand/Sequence	Identifier Link
-300MOTIFZMZEIN	153 (-) RTGAGTCAT	S000002
AACACOREOSGLUB1	356 (-) AACAAAC	S000353
ABRELATERD1	641 (-) ACGTG	S000414
ABRELATERD1	1000 (-) ACGTG	S000414
ABRERATCAL	999 (-) MACGYGB	S000507
ACGTATERD1	127 (+) ACGT	S000415
ACGTATERD1	642 (+) ACGT	S000415
ACGTATERD1	1001 (+) ACGT	S000415
ACGTATERD1	127 (-) ACGT	S000415
ACGTATERD1	642 (-) ACGT	S000415
ACGTATERD1	1001 (-) ACGT	S000415
AMYBOX1	961 (-) TAACARA	S000020
AMYBOX2	288 (+) TATCCAT	S000021
AMYBOX2	776 (-) TATCCAT	S000021
ANAERO1CONSENSUS	524 (+) AAACAAA	S000477
ANAERO1CONSENSUS	357 (-) AAACAAA	S000477
ANAERO2CONSENSUS	1316 (+) AGCAGC	S000478
ANAERO3CONSENSUS	920 (+) TCATCAC	S000479

ARR1AT	564 (+) NGATT	S000454
ARR1AT	1438 (+) NGATT	S000454
ARR1AT	42 (+) NGATT	S000454
ARR1AT	928 (+) NGATT	S000454
ARR1AT	1240 (+) NGATT	S000454
ARR1AT	1443 (+) NGATT	S000454
ARR1AT	190 (-) NGATT	S000454
ARR1AT	611 (-) NGATT	S000454
ARR1AT	816 (-) NGATT	S000454
ARR1AT	852 (-) NGATT	S000454
ARR1AT	1186 (-) NGATT	S000454
ARR1AT	1264 (-) NGATT	S000454
ASF1MOTIFCAMV	125 (+) TGACG	S000024
BIHD1OS	800 (+) TGTCA	S000498
BIHD1OS	1151 (+) TGTCA	S000498
BIHD1OS	90 (-) TGTCA	S000498
BIHD1OS	183 (-) TGTCA	S000498
BIHD1OS	744 (-) TGTCA	S000498
BIHD1OS	1168 (-) TGTCA	S000498
BOXIINTPATPB	54 (+) ATAGAA	S000296
BOXIINTPATPB	347 (-) ATAGAA	S000296
BOXIINTPATPB	1223 (-) ATAGAA	S000296
BOXIINTPATPB	1234 (-) ATAGAA	S000296
CAATBOX1	400 (+) CAAT	S000028
CAATBOX1	426 (+) CAAT	S000028
CAATBOX1	618 (+) CAAT	S000028
CAATBOX1	815 (+) CAAT	S000028
CAATBOX1	1050 (+) CAAT	S000028
CAATBOX1	1062 (+) CAAT	S000028
CAATBOX1	1114 (+) CAAT	S000028
CAATBOX1	1279 (+) CAAT	S000028
CAATBOX1	27 (-) CAAT	S000028
CAATBOX1	88 (-) CAAT	S000028
CAATBOX1	195 (-) CAAT	S000028
CAATBOX1	241 (-) CAAT	S000028
CAATBOX1	316 (-) CAAT	S000028
CAATBOX1	798 (-) CAAT	S000028
CAATBOX1	1142 (-) CAAT	S000028
CAATBOX1	1238 (-) CAAT	S000028
CAATBOX1	1293 (-) CAAT	S000028
CAATBOX1	1309 (-) CAAT	S000028
CAATBOX1	1334 (-) CAAT	S000028
CACTFTPPCA1	114 (+) YACT	S000449
CACTFTPPCA1	803 (+) YACT	S000449
CACTFTPPCA1	958 (+) YACT	S000449
CACTFTPPCA1	1213 (+) YACT	S000449
CACTFTPPCA1	1219 (+) YACT	S000449
CACTFTPPCA1	668 (+) YACT	S000449
CACTFTPPCA1	828 (+) YACT	S000449
CACTFTPPCA1	1203 (+) YACT	S000449
CACTFTPPCA1	131 (-) YACT	S000449
CACTFTPPCA1	207 (-) YACT	S000449
CACTFTPPCA1	541 (-) YACT	S000449
CACTFTPPCA1	717 (-) YACT	S000449

CACTFTPPCA1	788	(-) YACT	S000449
CARGCW8GAT	93	(+) CWWWWWWWWG	S000431
CARGCW8GAT	171	(+) CWWWWWWWWG	S000431
CARGCW8GAT	93	(-) CWWWWWWWWG	S000431
CARGCW8GAT	171	(-) CWWWWWWWWG	S000431
CBFHV	1178	(+) RYCGAC	S000497
CCAATBOX1	1334	(-) CCAAT	S000030
CIACADIANLELHC	609	(+) CAANNNNATC	S000252
CIACADIANLELHC	1062	(+) CAANNNNATC	S000252
CIACADIANLELHC	1093	(+) CAANNNNATC	S000252
CURECORECR	760	(+) GTAC	S000493
CURECORECR	995	(+) GTAC	S000493
CURECORECR	1357	(+) GTAC	S000493
CURECORECR	760	(-) GTAC	S000493
CURECORECR	995	(-) GTAC	S000493
CURECORECR	1357	(-) GTAC	S000493
DOFCOREZM	272	(+) AAAG	S000265
DOFCOREZM	330	(+) AAAG	S000265
DOFCOREZM	460	(+) AAAG	S000265
DOFCOREZM	598	(+) AAAG	S000265
DOFCOREZM	861	(+) AAAG	S000265
DOFCOREZM	37	(-) AAAG	S000265
DOFCOREZM	345	(-) AAAG	S000265
DOFCOREZM	374	(-) AAAG	S000265
DOFCOREZM	394	(-) AAAG	S000265
DOFCOREZM	754	(-) AAAG	S000265
DOFCOREZM	805	(-) AAAG	S000265
DOFCOREZM	913	(-) AAAG	S000265
DOFCOREZM	960	(-) AAAG	S000265
DOFCOREZM	1221	(-) AAAG	S000265
DOFCOREZM	1232	(-) AAAG	S000265
DOFCOREZM	1284	(-) AAAG	S000265
DPBFCOREDCDC3	514	(+) ACACNNG	S000292
DPBFCOREDCDC3	638	(+) ACACNNG	S000292
DPBFCOREDCDC3	997	(+) ACACNNG	S000292
DRE2COREZMRAB17	1178	(+) ACCGAC	S000402
DRECRTCOREAT	1178	(+) RCCGAC	S000418
E2FCONSENSUS	735	(-) WTTSSCSS	S000476
EBOXBNNAPA	924	(+) CANNTG	S000144
EBOXBNNAPA	924	(-) CANNTG	S000144
EECCRCAH1	601	(+) GANTTNC	S000494
EECCRCAH1	473	(-) GANTTNC	S000494
GAREAT	961	(-) TAACAAR	S000439
GATABOX	286	(+) GATA	S000039
GATABOX	481	(+) GATA	S000039
GATABOX	765	(+) GATA	S000039
GATABOX	779	(+) GATA	S000039
GATABOX	80	(-) GATA	S000039
GATABOX	161	(-) GATA	S000039
GATABOX	288	(-) GATA	S000039
GATABOX	350	(-) GATA	S000039
GATABOX	615	(-) GATA	S000039
GATABOX	790	(-) GATA	S000039
GATABOX	846	(-) GATA	S000039

GATABOX	1099 (-) GATA	S000039
GCN4OSGLUB1	154 (-) TGAGTCA	S000277
GLMHVCHORD	153 (+) RTGASTCAT	S000451
GLMHVCHORD	153 (-) RTGASTCAT	S000451
GT1CONSENSUS	57 (+) GRWAAW	S000198
GT1CONSENSUS	225 (+) GRWAAW	S000198
GT1CONSENSUS	254 (+) GRWAAW	S000198
GT1CONSENSUS	255 (+) GRWAAW	S000198
GT1CONSENSUS	297 (+) GRWAAW	S000198
GT1CONSENSUS	333 (+) GRWAAW	S000198
GT1CONSENSUS	334 (+) GRWAAW	S000198
GT1CONSENSUS	463 (+) GRWAAW	S000198
GT1CONSENSUS	864 (+) GRWAAW	S000198
GT1CONSENSUS	865 (+) GRWAAW	S000198
GT1CONSENSUS	1393 (+) GRWAAW	S000198
GT1CONSENSUS	389 (-) GRWAAW	S000198
GT1CONSENSUS	836 (-) GRWAAW	S000198
GT1CONSENSUS	948 (-) GRWAAW	S000198
GT1CONSENSUS	1019 (-) GRWAAW	S000198
GT1CONSENSUS	1035 (-) GRWAAW	S000198
GT1CONSENSUS	1227 (-) GRWAAW	S000198
GT1CONSENSUS	305 (-) GRWAAW	S000198
GT1CONSENSUS	368 (-) GRWAAW	S000198
GT1CONSENSUS	369 (-) GRWAAW	S000198
GT1CONSENSUS	710 (-) GRWAAW	S000198
GT1CONSENSUS	1097 (-) GRWAAW	S000198
GT1CONSENSUS	1386 (-) GRWAAW	S000198
GT1CONSENSUS	1404 (-) GRWAAW	S000198
GT1GMSCAM4	225 (+) GAAAAA	S000453
GT1GMSCAM4	255 (+) GAAAAA	S000453
GT1GMSCAM4	297 (+) GAAAAA	S000453
GT1GMSCAM4	463 (+) GAAAAA	S000453
GT1GMSCAM4	368 (-) GAAAAA	S000453
GT1GMSCAM4	1386 (-) GAAAAA	S000453
GT1GMSCAM4	1404 (-) GAAAAA	S000453
GTGANTG10	284 (+) GTGA	S000378
GTGANTG10	561 (+) GTGA	S000378
GTGANTG10	989 (+) GTGA	S000378
GTGANTG10	113 (-) GTGA	S000378
GTGANTG10	397 (-) GTGA	S000378
GTGANTG10	802 (-) GTGA	S000378
GTGANTG10	923 (-) GTGA	S000378
GTGANTG10	932 (-) GTGA	S000378
GTGANTG10	957 (-) GTGA	S000378
GTGANTG10	1101 (-) GTGA	S000378
GTGANTG10	1162 (-) GTGA	S000378
GTGANTG10	1210 (-) GTGA	S000378
GTGANTG10	1216 (-) GTGA	S000378
HEXMOTIFTAH3H4	125 (-) ACGTCA	S000053
IBOX	765 (+) GATAAG	S000124
IBOXCORE	765 (+) GATAA	S000199
IBOXCORE	1098 (-) GATAA	S000199
INRNTPSADB	938 (+) YTCANTYY	S000395
INRNTPSADB	1032 (+) YTCANTYY	S000395

INRNTPSADB	1277	(+) YTCANTYY	S000395
INRNTPSADB	1407	(+) YTCANTYY	S000395
INRNTPSADB	324	(-) YTCANTYY	S000395
INRNTPSADB	449	(-) YTCANTYY	S000395
INRNTPSADB	468	(-) YTCANTYY	S000395
INRNTPSADB	211	(-) YTCANTYY	S000395
L1BOXATPDF1	66	(+) TAAATGYA	S000386
LTRE1HVBLT49	950	(-) CCGAAA	S000250
LTRECOREATCOR15	1179	(+) CCGAC	S000153
MARABOX1	60	(+) AATAAAYAAA	S000063
MARABOX1	521	(+) AATAAAYAAA	S000063
MARABOX1	868	(+) AATAAAYAAA	S000063
MARTBOX	363	(+) TTWTWTTWTT	S000067
MARTBOX	591	(-) TTWTWTTWTT	S000067
MARTBOX	872	(-) TTWTWTTWTT	S000067
MARTBOX	875	(-) TTWTWTTWTT	S000067
MARTBOX	877	(-) TTWTWTTWTT	S000067
MYB1AT	1250	(+) WAACCA	S000408
MYB2AT	650	(+) TAACTG	S000177
MYB2CONSENSUSAT	650	(+) YAACKG	S000409
MYBCORE	148	(+) CNGTTR	S000176
MYBCORE	1189	(-) CNGTTR	S000176
MYBCORE	650	(-) CNGTTR	S000176
MYBGAHV	961	(-) TAACAAA	S000181
MYBPLANT	1102	(+) MACCWAMC	S000167
MYBST1	778	(+) GGATA	S000180
MYBST1	161	(-) GGATA	S000180
MYBST1	288	(-) GGATA	S000180
MYBST1	790	(-) GGATA	S000180
MYCATERD1	924	(-) CATGTG	S000413
MYCATRD22	924	(+) CACATG	S000174
MYCCONSENSUSAT	924	(+) CANNTG	S000407
MYCCONSENSUSAT	924	(-) CANNTG	S000407
NODCON2GM	1131	(+) CTCTT	S000462
NODCON2GM	538	(-) CTCTT	S000462
NODCON2GM	1297	(-) CTCTT	S000462
OSE2ROOTNODULE	1131	(+) CTCTT	S000468
OSE2ROOTNODULE	538	(-) CTCTT	S000468
OSE2ROOTNODULE	1297	(-) CTCTT	S000468
PE2FNTRNR1A	735	(-) ATTCGCGC	S000455
POLASIG1	60	(+) AATAAA	S000080
POLASIG1	64	(+) AATAAA	S000080
POLASIG1	321	(+) AATAAA	S000080
POLASIG1	521	(+) AATAAA	S000080
POLASIG1	662	(+) AATAAA	S000080
POLASIG1	868	(+) AATAAA	S000080
POLASIG1	872	(+) AATAAA	S000080
POLASIG1	880	(+) AATAAA	S000080
POLASIG1	199	(-) AATAAA	S000080
POLASIG1	364	(-) AATAAA	S000080
POLASIG1	386	(-) AATAAA	S000080
POLASIG1	833	(-) AATAAA	S000080
POLASIG1	915	(-) AATAAA	S000080
POLASIG1	1412	(-) AATAAA	S000080

POLASIG2	455 (+) AATTAAA	S000081
POLASIG2	619 (+) AATTAAA	S000081
POLASIG2	678 (+) AATTAAA	S000081
POLASIG2	1427 (+) AATTAAA	S000081
POLASIG2	379 (-) AATTAAA	S000081
POLASIG3	11 (+) AATAAT	S000088
POLASIG3	337 (+) AATAAT	S000088
POLASIG3	975 (+) AATAAT	S000088
POLASIG3	1051 (+) AATAAT	S000088
POLASIG3	24 (-) AATAAT	S000088
POLASIG3	701 (-) AATAAT	S000088
POLASIG3	704 (-) AATAAT	S000088
POLASIG3	1242 (-) AATAAT	S000088
POLLEN1LELAT52	56 (+) AGAAA	S000245
POLLEN1LELAT52	210 (+) AGAAA	S000245
POLLEN1LELAT52	224 (+) AGAAA	S000245
POLLEN1LELAT52	269 (+) AGAAA	S000245
POLLEN1LELAT52	296 (+) AGAAA	S000245
POLLEN1LELAT52	462 (+) AGAAA	S000245
POLLEN1LELAT52	693 (+) AGAAA	S000245
POLLEN1LELAT52	120 (-) AGAAA	S000245
POLLEN1LELAT52	137 (-) AGAAA	S000245
POLLEN1LELAT52	342 (-) AGAAA	S000245
POLLEN1LELAT52	346 (-) AGAAA	S000245
POLLEN1LELAT52	391 (-) AGAAA	S000245
POLLEN1LELAT52	806 (-) AGAAA	S000245
POLLEN1LELAT52	838 (-) AGAAA	S000245
POLLEN1LELAT52	942 (-) AGAAA	S000245
POLLEN1LELAT52	1021 (-) AGAAA	S000245
POLLEN1LELAT52	1222 (-) AGAAA	S000245
POLLEN1LELAT52	1233 (-) AGAAA	S000245
POLLEN1LELAT52	1388 (-) AGAAA	S000245
PREATPRODH	156 (+) ACTCAT	S000450
PREATPRODH	1303 (+) ACTCAT	S000450
PRECONSCRHSP70A	1179 (+) SCGAYNRNNNNNNNNNNNNNNHD	S000506
PYRIMIDINEBOXOSRAMY1A	373 (+) CCTTTT	S000259
PYRIMIDINEBOXOSRAMY1A	753 (+) CCTTTT	S000259
PYRIMIDINEBOXOSRAMY1A	860 (-) CCTTTT	S000259
QARBNEXTA	999 (-) AACGTGT	S000244
RAV1AAT	48 (+) CAACA	S000314
RAV1AAT	1047 (+) CAACA	S000314
RAV1AAT	1111 (+) CAACA	S000314
RAV1AAT	888 (-) CAACA	S000314
RHERPATEXPA7	1162 (+) KCACGW	S000512
ROOTMOTIFTAPOX1	278 (+) ATATT	S000098
ROOTMOTIFTAPOX1	585 (+) ATATT	S000098
ROOTMOTIFTAPOX1	780 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1307 (+) ATATT	S000098
ROOTMOTIFTAPOX1	14 (-) ATATT	S000098
ROOTMOTIFTAPOX1	401 (-) ATATT	S000098
ROOTMOTIFTAPOX1	822 (-) ATATT	S000098
ROOTMOTIFTAPOX1	844 (-) ATATT	S000098
RYREPEATBNNAPA	70 (-) CATGCA	S000264

RYREPEATGMGY2	69	(-)	CATGCAT	S000105
RYREPEATLEGUMINBOX	69	(-)	CATGCAY	S000100
S1FBOXSORPS1L21	1116	(+)	ATGGTA	S000223
SEBFCONSSTPR10A	799	(+)	YTGTCWC	S000391
SEF1MOTIF	585	(+)	ATATTTAWW	S000006
SEF3MOTIFGM	1107	(+)	AACCCA	S000115
SEF4MOTIFGM7S	383	(+)	RTTTTTR	S000103
SEF4MOTIFGM7S	707	(+)	RTTTTTR	S000103
SEF4MOTIFGM7S	782	(+)	RTTTTTR	S000103
SEF4MOTIFGM7S	655	(+)	RTTTTTR	S000103
SEF4MOTIFGM7S	1370	(+)	RTTTTTR	S000103
SEF4MOTIFGM7S	747	(-)	RTTTTTR	S000103
SEF4MOTIFGM7S	719	(-)	RTTTTTR	S000103
SEF4MOTIFGM7S	882	(-)	RTTTTTR	S000103
SORLREP3AT	502	(-)	TGTATATAT	S000488
SP8BFIBSP8BIB	828	(+)	TACTATT	S000184
SURECOREATSULTR11	216	(+)	GAGAC	S000499
SURECOREATSULTR11	644	(-)	GAGAC	S000499
T/GBOXATPIN2	1000	(-)	AACGTG	S000458
TATABOX2	971	(+)	TATAAAT	S000109
TATABOX5	117	(+)	TTATTT	S000203
TATABOX5	200	(+)	TTATTT	S000203
TATABOX5	365	(+)	TTATTT	S000203
TATABOX5	387	(+)	TTATTT	S000203
TATABOX5	705	(+)	TTATTT	S000203
TATABOX5	834	(+)	TTATTT	S000203
TATABOX5	946	(+)	TTATTT	S000203
TATABOX5	1413	(+)	TTATTT	S000203
TATABOX5	10	(-)	TTATTT	S000203
TATABOX5	59	(-)	TTATTT	S000203
TATABOX5	63	(-)	TTATTT	S000203
TATABOX5	320	(-)	TTATTT	S000203
TATABOX5	336	(-)	TTATTT	S000203
TATABOX5	867	(-)	TTATTT	S000203
TATABOX5	871	(-)	TTATTT	S000203
TATABOX5	879	(-)	TTATTT	S000203
TATABOX5	974	(-)	TTATTT	S000203
TATABOXOSPAL	586	(+)	TATTTAA	S000400
TATABOXOSPAL	8	(-)	TATTTAA	S000400
TATCCAOSAMY	288	(+)	TATCCA	S000403
TATCCAOSAMY	790	(+)	TATCCA	S000403
TATCCAOSAMY	777	(-)	TATCCA	S000403
TATCCAYMOTIFOSRAMY3D	288	(+)	TATCCAY	S000256
TATCCAYMOTIFOSRAMY3D	776	(-)	TATCCAY	S000256
TBOXATGAPB	959	(+)	ACTTTG	S000383
TGACGTVMAMY	125	(+)	TGACGT	S000377
TRANSINITDICOTS	1340	(-)	AMNAUGGC	S000201
TRANSINITMONOCOTS	1340	(-)	RMNAUGGC	S000202
WBOXATNPR1	89	(+)	TTGAC	S000390
WBOXATNPR1	124	(+)	TTGAC	S000390
WBOXATNPR1	743	(+)	TTGAC	S000390
WBOXHVISO1	154	(+)	TGACT	S000442
WBOXNTERF3	154	(+)	TGACY	S000457
WBOXNTERF3	955	(-)	TGACY	S000457

WRKY71OS	90 (+) TGAC	S000447
WRKY71OS	125 (+) TGAC	S000447
WRKY71OS	154 (+) TGAC	S000447
WRKY71OS	183 (+) TGAC	S000447
WRKY71OS	744 (+) TGAC	S000447
WRKY71OS	1168 (+) TGAC	S000447
WRKY71OS	801 (-) TGAC	S000447
WRKY71OS	956 (-) TGAC	S000447
WRKY71OS	1152 (-) TGAC	S000447

Segment of chromosome 1 upstream of LJFgene1 start codon:

TTTCTGCATTACAGATAGCTACCATTTTCAGTGCATGGAAAACTTGGAAA
TTACATGACAACAAAATCAGTAATGAAACCAATAAAAACCAAGTCAACAT
GATCCATTGCATAGTAAAGACTAAAATGAGTGAGTACCTTCAGTTTCAAA
TAAGACTATTCCCACTTCATCATATTTGCTATAGATCAGCTGCACATAAA
ACAAAACAAAAACAAAAGGATCATGTTATTGAAGCTATCAACCAGCAAAA
CCAAACACACAAGATTGACATAAGAAAAACCAAAAAAAAAAAAAAAAATGA
AACTTTGAATAGAAAAGAGTGTGGGAATGATTACTACCTTTTCCTGTACT
AGCATGGAGCAGACTTTCTCAATTTCAGGAAGAACAGAATGCATGGATGG
GCCCACATCAAGCAGCAAAAGCAAGGCATTTTTTTTTCCAATAATTTAAT
TGTGGCACCAAAAGACAAACACAAATTTACGCATTCAGAAGTAACGAAAT
TTACAATAATAAAAAGAAACTAAAACTGTTCCTAACATGATTCTATAGCG
AAAGACCAAATTTTGAGAACGAGTCAGCAACATAATTAAAAAGAAAGATG
GTAACATTCTAAATGGTGTTTCCCGGGAAAATGAGTAAAGTTGAAGCAAA
ACCCTAAACCTAGATAGCTGAAGCTGGGGTCAAAAGAAAAAATGGATGCT
TCAAATGAAGAGCACGCTTTTCTTTCATGGACGGTGATATATCGGTGCAT
TTTTATTTTTATTTTTCAGTTGCAAAGGAACTTTGGAGCGAAATAAAGTT
TGTAAAAGCATCGAAACCTTTCCTTTCTGTGGAACTGAATTTGTGAATCA
TTAGATTAATTACGAAATTCAAACACTGCATTTTAAACTAAAATCTAAAC
ATCACTTAAAACATCCACTTTGTCCTCGTATAAAAGAAATTATCATAAGC
CAAAACATACTAAGGATAAATCCTAGGCCCGAAGAAAACATCGTAATAAC
ATTCTTTTTACTTTATTTCATCTCTATTTTTGTCTTCTAGTAGTATTTCC
CCCCCTCCTTATTTCACTCCTAATCAAAACAAAATAGAGAAAAAATGAAA
CTCGGATGTATATGCGAATAGACATAATTTTCTAAATACATACTAATTAA
AATTAAAAAATGTAAGCATTATTATTTTTACCAATTTGACAAAAACCTTT
TCGTACAGAGATAAGCATTTGTGGATATTTTTAGTATCCAAAATTGTCCA
TTTTCTAAAAAAATCGAAATATATATTTATTTTCTAAATATCCAAATCCA
TAAAAAGGAGAAAGAAAAAACAAAAAAAAAATGTTGCGGAAACGAAGCGT
CCCACCACCCACCTTCTTTTTATATTCATCATTACATGATTCACATCTCA
TTCCATATTTTCGGGTCACATTCTCAAATTATTATAACTAATTTCGTCAT
GTGAAGATACGTTCATAAGTTCCTTAATTTTCTTGAACCTTTATTTTCAG
CTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCT

Element name	Site/Strand/Sequence	Identifier link
-300ELEMENT	771 (+) TGHAAARK	S000122
-300ELEMENT	801 (+) TGHAAARK	S000122
-300ELEMENT	761 (-) TGHAAARK	S000122
2SSEEDPROTBANAPA	252 (+) CAAACAC	S000143
2SSEEDPROTBANAPA	466 (+) CAAACAC	S000143
2SSEEDPROTBANAPA	870 (+) CAAACAC	S000143
5659BOXLELAT5659	837 (+) GAAWTTGTGA	S000280
ACGTATERD1	1459 (+) ACGT	S000415
ACGTATERD1	1459 (-) ACGT	S000415

ANAERO1CONSENSUS	199 (+) AAACAAA	S000477
ANAERO1CONSENSUS	204 (+) AAACAAA	S000477
ANAERO1CONSENSUS	210 (+) AAACAAA	S000477
ANAERO1CONSENSUS	1077 (+) AAACAAA	S000477
ANAERO1CONSENSUS	1318 (+) AAACAAA	S000477
ANAERO2CONSENSUS	411 (+) AGCAGC	S000478
ARR1AT	262 (+) NGATT	S000454
ARR1AT	853 (+) NGATT	S000454
ARR1AT	328 (+) NGATT	S000454
ARR1AT	538 (+) NGATT	S000454
ARR1AT	1387 (+) NGATT	S000454
ARR1AT	65 (-) NGATT	S000454
ARR1AT	846 (-) NGATT	S000454
ARR1AT	892 (-) NGATT	S000454
ARR1AT	969 (-) NGATT	S000454
ARR1AT	1072 (-) NGATT	S000454
ARR1AT	1262 (-) NGATT	S000454
ARR1AT	1295 (-) NGATT	S000454
ASF1MOTIFCAMV	1445 (-) TGACG	S000024
BIHD1OS	56 (-) TGTCA	S000498
BIHD1OS	266 (-) TGTCA	S000498
BIHD1OS	1187 (-) TGTCA	S000498
BOXIINTPATPB	309 (+) ATAGAA	S000296
BOXIINTPATPB	541 (-) ATAGAA	S000296
CAATBOX1	80 (+) CAAT	S000028
CAATBOX1	370 (+) CAAT	S000028
CAATBOX1	439 (+) CAAT	S000028
CAATBOX1	504 (+) CAAT	S000028
CAATBOX1	1182 (+) CAAT	S000028
CAATBOX1	1508 (+) CAAT	S000028
CAATBOX1	1520 (+) CAAT	S000028
CAATBOX1	106 (-) CAAT	S000028
CAATBOX1	228 (-) CAAT	S000028
CAATBOX1	264 (-) CAAT	S000028
CAATBOX1	449 (-) CAAT	S000028
CAATBOX1	1243 (-) CAAT	S000028
CACTFTPPCA1	163 (+) YACT	S000449
CACTFTPPCA1	874 (+) YACT	S000449
CACTFTPPCA1	903 (+) YACT	S000449
CACTFTPPCA1	916 (+) YACT	S000449
CACTFTPPCA1	1065 (+) YACT	S000449
CACTFTPPCA1	332 (+) YACT	S000449
CACTFTPPCA1	347 (+) YACT	S000449
CACTFTPPCA1	958 (+) YACT	S000449
CACTFTPPCA1	1009 (+) YACT	S000449
CACTFTPPCA1	1141 (+) YACT	S000449
CACTFTPPCA1	30 (-) YACT	S000449
CACTFTPPCA1	69 (-) YACT	S000449
CACTFTPPCA1	113 (-) YACT	S000449
CACTFTPPCA1	129 (-) YACT	S000449
CACTFTPPCA1	133 (-) YACT	S000449
CACTFTPPCA1	318 (-) YACT	S000449
CACTFTPPCA1	490 (-) YACT	S000449
CACTFTPPCA1	634 (-) YACT	S000449

CACTFTPPCA1	1039 (-) YACT	S000449
CACTFTPPCA1	1042 (-) YACT	S000449
CACTFTPPCA1	1233 (-) YACT	S000449
CANBNNAPA	252 (+) CNAACAC	S000148
CANBNNAPA	466 (+) CNAACAC	S000148
CANBNNAPA	870 (+) CNAACAC	S000148
CCAATBOX1	79 (+) CCAAT	S000030
CCAATBOX1	438 (+) CCAAT	S000030
CCAATBOX1	1181 (+) CCAAT	S000030
CIACADIANLELHC	59 (+) CAANNNNATC	S000252
CIACADIANLELHC	95 (+) CAANNNNATC	S000252
CIACADIANLELHC	213 (+) CAANNNNATC	S000252
CIACADIANLELHC	1520 (+) CAANNNNATC	S000252
CMSRE1IBSPOA	728 (+) TGGACGG	S000511
CURECORECR	134 (+) GTAC	S000493
CURECORECR	346 (+) GTAC	S000493
CURECORECR	1203 (+) GTAC	S000493
CURECORECR	134 (-) GTAC	S000493
CURECORECR	346 (-) GTAC	S000493
CURECORECR	1203 (-) GTAC	S000493
DOFCOREZM	116 (+) AAAG	S000265
DOFCOREZM	215 (+) AAAG	S000265
DOFCOREZM	314 (+) AAAG	S000265
DOFCOREZM	418 (+) AAAG	S000265
DOFCOREZM	461 (+) AAAG	S000265
DOFCOREZM	513 (+) AAAG	S000265
DOFCOREZM	551 (+) AAAG	S000265
DOFCOREZM	590 (+) AAAG	S000265
DOFCOREZM	594 (+) AAAG	S000265
DOFCOREZM	637 (+) AAAG	S000265
DOFCOREZM	683 (+) AAAG	S000265
DOFCOREZM	774 (+) AAAG	S000265
DOFCOREZM	795 (+) AAAG	S000265
DOFCOREZM	805 (+) AAAG	S000265
DOFCOREZM	933 (+) AAAG	S000265
DOFCOREZM	1304 (+) AAAG	S000265
DOFCOREZM	1311 (+) AAAG	S000265
DOFCOREZM	303 (-) AAAG	S000265
DOFCOREZM	338 (-) AAAG	S000265
DOFCOREZM	364 (-) AAAG	S000265
DOFCOREZM	717 (-) AAAG	S000265
DOFCOREZM	722 (-) AAAG	S000265
DOFCOREZM	781 (-) AAAG	S000265
DOFCOREZM	818 (-) AAAG	S000265
DOFCOREZM	823 (-) AAAG	S000265
DOFCOREZM	918 (-) AAAG	S000265
DOFCOREZM	1004 (-) AAAG	S000265
DOFCOREZM	1011 (-) AAAG	S000265
DOFCOREZM	1197 (-) AAAG	S000265
DOFCOREZM	1366 (-) AAAG	S000265
DOFCOREZM	1489 (-) AAAG	S000265
DPBFCOREDCDC3	257 (+) ACACNNG	S000292
E2FCONSENSUS	619 (+) WTTSSCSS	S000476
E2FCONSENSUS	1046 (+) WTTSSCSS	S000476

E2FCONSENSUS	623 (-) WTTSSCSS	S000476
EBOXBNNAPA	187 (+) CANNTG	S000144
EBOXBNNAPA	702 (+) CANNTG	S000144
EBOXBNNAPA	767 (+) CANNTG	S000144
EBOXBNNAPA	1216 (+) CANNTG	S000144
EBOXBNNAPA	1448 (+) CANNTG	S000144
EBOXBNNAPA	187 (-) CANNTG	S000144
EBOXBNNAPA	702 (-) CANNTG	S000144
EBOXBNNAPA	767 (-) CANNTG	S000144
EBOXBNNAPA	1216 (-) CANNTG	S000144
EBOXBNNAPA	1448 (-) CANNTG	S000144
EECCRCAH1	362 (+) GANTTNC	S000494
EECCRCAH1	864 (-) GANTTNC	S000494
EECCRCAH1	1097 (-) GANTTNC	S000494
ELRECOREPCRP1	678 (-) TTGACC	S000142
ERELEE4	866 (+) AWTTCAAA	S000037
GATABOX	14 (+) GATA	S000039
GATABOX	663 (+) GATA	S000039
GATABOX	736 (+) GATA	S000039
GATABOX	965 (+) GATA	S000039
GATABOX	1210 (+) GATA	S000039
GATABOX	1224 (+) GATA	S000039
GATABOX	1456 (+) GATA	S000039
GATABOX	236 (-) GATA	S000039
GATABOX	740 (-) GATA	S000039
GATABOX	941 (-) GATA	S000039
GATABOX	1235 (-) GATA	S000039
GATABOX	1289 (-) GATA	S000039
GT1CONSENSUS	37 (+) GRWAAW	S000198
GT1CONSENSUS	46 (+) GRWAAW	S000198
GT1CONSENSUS	274 (+) GRWAAW	S000198
GT1CONSENSUS	626 (+) GRWAAW	S000198
GT1CONSENSUS	627 (+) GRWAAW	S000198
GT1CONSENSUS	686 (+) GRWAAW	S000198
GT1CONSENSUS	965 (+) GRWAAW	S000198
GT1CONSENSUS	1089 (+) GRWAAW	S000198
GT1CONSENSUS	1314 (+) GRWAAW	S000198
GT1CONSENSUS	24 (-) GRWAAW	S000198
GT1CONSENSUS	939 (-) GRWAAW	S000198
GT1CONSENSUS	1045 (-) GRWAAW	S000198
GT1CONSENSUS	1127 (-) GRWAAW	S000198
GT1CONSENSUS	1250 (-) GRWAAW	S000198
GT1CONSENSUS	1279 (-) GRWAAW	S000198
GT1CONSENSUS	1407 (-) GRWAAW	S000198
GT1CONSENSUS	1477 (-) GRWAAW	S000198
GT1CONSENSUS	1493 (-) GRWAAW	S000198
GT1CONSENSUS	339 (-) GRWAAW	S000198
GT1CONSENSUS	433 (-) GRWAAW	S000198
GT1CONSENSUS	434 (-) GRWAAW	S000198
GT1CONSENSUS	762 (-) GRWAAW	S000198
GT1CONSENSUS	1177 (-) GRWAAW	S000198
GT1GMSCAM4	274 (+) GAAAAA	S000453
GT1GMSCAM4	686 (+) GAAAAA	S000453
GT1GMSCAM4	1089 (+) GAAAAA	S000453

GT1GMSCAM4	1314 (+) GAAAAA	S000453
GT1GMSCAM4	433 (-) GAAAAA	S000453
GT1GMSCAM4	762 (-) GAAAAA	S000453
GTGANTG10	130 (+) GTGA	S000378
GTGANTG10	734 (+) GTGA	S000378
GTGANTG10	843 (+) GTGA	S000378
GTGANTG10	1451 (+) GTGA	S000378
GTGANTG10	902 (-) GTGA	S000378
GTGANTG10	1064 (-) GTGA	S000378
GTGANTG10	1391 (-) GTGA	S000378
GTGANTG10	1416 (-) GTGA	S000378
IBOX	1210 (+) GATAAG	S000124
IBOXCORE	965 (+) GATAA	S000199
IBOXCORE	1210 (+) GATAA	S000199
IBOXCORE	940 (-) GATAA	S000199
INRNTPSADB	368 (+) YTCANTYY	S000395
INRNTPSADB	1397 (+) YTCANTYY	S000395
INRNTPSADB	139 (+) YTCANTYY	S000395
INRNTPSADB	1063 (+) YTCANTYY	S000395
INRNTPSADB	123 (-) YTCANTYY	S000395
INRNTPSADB	294 (-) YTCANTYY	S000395
INRNTPSADB	628 (-) YTCANTYY	S000395
INRNTPSADB	1092 (-) YTCANTYY	S000395
INRNTPSADB	832 (-) YTCANTYY	S000395
LTRE1HVBLT49	1409 (-) CCGAAA	S000250
MARTBOX	751 (+) TTWTWTTWTT	S000067
MARTBOX	757 (+) TTWTWTTWTT	S000067
MARTBOX	282 (-) TTWTWTTWTT	S000067
MARTBOX	283 (-) TTWTWTTWTT	S000067
MARTBOX	284 (-) TTWTWTTWTT	S000067
MARTBOX	285 (-) TTWTWTTWTT	S000067
MARTBOX	286 (-) TTWTWTTWTT	S000067
MARTBOX	287 (-) TTWTWTTWTT	S000067
MARTBOX	288 (-) TTWTWTTWTT	S000067
MARTBOX	505 (-) TTWTWTTWTT	S000067
MARTBOX	1322 (-) TTWTWTTWTT	S000067
MYB1AT	76 (+) WAACCA	S000408
MYB1AT	86 (+) WAACCA	S000408
MYB1AT	248 (+) WAACCA	S000408
MYB1AT	277 (+) WAACCA	S000408
MYB2CONSENSUSAT	767 (-) YAACKG	S000409
MYBCORE	767 (+) CNGTTR	S000176
MYBPLANT	249 (+) MACCWAMC	S000167
MYBST1	964 (+) GGATA	S000180
MYBST1	1223 (+) GGATA	S000180
MYBST1	1235 (-) GGATA	S000180
MYBST1	1289 (-) GGATA	S000180
MYCATERD1	1448 (+) CATGTG	S000413
MYCATRD22	1448 (-) CACATG	S000174
MYCCONSENSUSAT	187 (+) CANNTG	S000407
MYCCONSENSUSAT	702 (+) CANNTG	S000407
MYCCONSENSUSAT	767 (+) CANNTG	S000407
MYCCONSENSUSAT	1216 (+) CANNTG	S000407
MYCCONSENSUSAT	1448 (+) CANNTG	S000407

MYCCONSENSUSAT	187 (-) CANNTG	S000407
MYCCONSENSUSAT	702 (-) CANNTG	S000407
MYCCONSENSUSAT	767 (-) CANNTG	S000407
MYCCONSENSUSAT	1216 (-) CANNTG	S000407
MYCCONSENSUSAT	1448 (-) CANNTG	S000407
NODCON1GM	594 (+) AAAGAT	S000461
NODCON2GM	315 (-) CTCTT	S000462
NODCON2GM	708 (-) CTCTT	S000462
NTBBF1ARROLB	1010 (+) ACTTTA	S000273
NTBBF1ARROLB	636 (-) ACTTTA	S000273
NTBBF1ARROLB	794 (-) ACTTTA	S000273
OSE1ROOTNODULE	594 (+) AAAGAT	S000467
OSE2ROOTNODULE	315 (-) CTCTT	S000468
OSE2ROOTNODULE	708 (-) CTCTT	S000468
P1BS	1108 (+) GNATATNC	S000459
P1BS	1108 (-) GNATATNC	S000459
PALBOXAPC	729 (-) CCGTCC	S000137
POLASIG1	81 (+) AATAAA	S000080
POLASIG1	508 (+) AATAAA	S000080
POLASIG1	792 (+) AATAAA	S000080
POLASIG1	752 (-) AATAAA	S000080
POLASIG1	758 (-) AATAAA	S000080
POLASIG1	1012 (-) AATAAA	S000080
POLASIG1	1276 (-) AATAAA	S000080
POLASIG1	1490 (-) AATAAA	S000080
POLASIG2	584 (+) AATTAAA	S000081
POLASIG2	1145 (+) AATTAAA	S000081
POLASIG2	1151 (+) AATTAAA	S000081
POLASIG2	445 (-) AATTAAA	S000081
POLASIG3	440 (+) AATAAT	S000088
POLASIG3	505 (+) AATAAT	S000088
POLASIG3	1509 (+) AATAAT	S000088
POLASIG3	1168 (-) AATAAT	S000088
POLASIG3	1171 (-) AATAAT	S000088
POLASIG3	1428 (-) AATAAT	S000088
POLLEN1LELAT52	273 (+) AGAAA	S000245
POLLEN1LELAT52	311 (+) AGAAA	S000245
POLLEN1LELAT52	515 (+) AGAAA	S000245
POLLEN1LELAT52	592 (+) AGAAA	S000245
POLLEN1LELAT52	685 (+) AGAAA	S000245
POLLEN1LELAT52	935 (+) AGAAA	S000245
POLLEN1LELAT52	983 (+) AGAAA	S000245
POLLEN1LELAT52	1088 (+) AGAAA	S000245
POLLEN1LELAT52	1309 (+) AGAAA	S000245
POLLEN1LELAT52	1313 (+) AGAAA	S000245
POLLEN1LELAT52	1 (-) AGAAA	S000245
POLLEN1LELAT52	365 (-) AGAAA	S000245
POLLEN1LELAT52	719 (-) AGAAA	S000245
POLLEN1LELAT52	824 (-) AGAAA	S000245
POLLEN1LELAT52	1129 (-) AGAAA	S000245
POLLEN1LELAT52	1252 (-) AGAAA	S000245
POLLEN1LELAT52	1281 (-) AGAAA	S000245
POLLEN1LELAT52	1479 (-) AGAAA	S000245
PREATPRODH	126 (-) ACTCAT	S000450

PREATPRODH	631 (-) ACTCAT	S000450
PROLAMINBOXOSGLUB1	771 (+) TGCAAAG	S000354
PYRIMIDINEBOXHVEPB1	432 (+) TTTTTTCC	S000298
PYRIMIDINEBOXOSRAMY1A	337 (+) CCTTTT	S000259
PYRIMIDINEBOXOSRAMY1A	1196 (+) CCTTTT	S000259
PYRIMIDINEBOXOSRAMY1A	214 (-) CCTTTT	S000259
PYRIMIDINEBOXOSRAMY1A	1303 (-) CCTTTT	S000259
RAV1AAT	59 (+) CAACA	S000314
RAV1AAT	95 (+) CAACA	S000314
RAV1AAT	578 (+) CAACA	S000314
RAV1AAT	1505 (+) CAACA	S000314
RAV1AAT	1332 (-) CAACA	S000314
REALPHALGLHCB21	77 (+) AACCAA	S000362
REALPHALGLHCB21	87 (+) AACCAA	S000362
REALPHALGLHCB21	249 (+) AACCAA	S000362
REALPHALGLHCB21	278 (+) AACCAA	S000362
ROOTMOTIFTAPOX1	172 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1225 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1273 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1372 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1405 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1268 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1287 (-) ATATT	S000098
RYREPEATBNNAPA	32 (-) CATGCA	S000264
RYREPEATBNNAPA	390 (-) CATGCA	S000264
RYREPEATGMGY2	389 (-) CATGCAT	S000105
RYREPEATLEGUMINBOX	389 (-) CATGCAY	S000100
RYREPEATLEGUMINBOX	31 (-) CATGCAY	S000100
S1FBOXSORPS1L21	598 (+) ATGGTA	S000223
S1FBOXSORPS1L21	20 (-) ATGGTA	S000223
SEF1MOTIF	1273 (+) ATATTTAWW	S000006
SEF4MOTIFGM7S	749 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	755 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	1026 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	1174 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	1227 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	207 (-) RTTTTTR	S000103
SEF4MOTIFGM7S	1190 (-) RTTTTTR	S000103
SEF4MOTIFGM7S	83 (-) RTTTTTR	S000103
SITEIIATCYTC	398 (+) TGGGCY	S000474
SITEIIATCYTC	400 (-) TGGGCY	S000474
SORLIP1AT	452 (-) GCCAC	S000482
SORLIP2AT	399 (+) GGGCC	S000483
SORLIP2AT	400 (-) GGGCC	S000483
SORLIP2AT	976 (-) GGGCC	S000483
SORLIP5AT	128 (+) GAGTGAG	S000486
SPHCOREZMC1	389 (-) TCCATGCAT	S000154
SREATMSD	964 (-) TTATCC	S000470
SURE1STPAT21	308 (+) AATAGAAAA	S000186
TAAAGSTKST1	115 (+) TAAAG	S000387
TAAAGSTKST1	636 (+) TAAAG	S000387
TAAAGSTKST1	794 (+) TAAAG	S000387
TAAAGSTKST1	1011 (-) TAAAG	S000387
TAAAGSTKST1	1489 (-) TAAAG	S000387

TATABOX5	753 (+) TTATTT	S000203
TATABOX5	759 (+) TTATTT	S000203
TATABOX5	1013 (+) TTATTT	S000203
TATABOX5	1059 (+) TTATTT	S000203
TATABOX5	1172 (+) TTATTT	S000203
TATABOX5	1277 (+) TTATTT	S000203
TATABOX5	1491 (+) TTATTT	S000203
TATABOX5	148 (-) TTATTT	S000203
TATABOX5	791 (-) TTATTT	S000203
TATCCACHVAL21	1221 (-) TATCCAC	S000416
TATCCAOSAMY	1235 (+) TATCCA	S000403
TATCCAOSAMY	1289 (+) TATCCA	S000403
TATCCAOSAMY	1222 (-) TATCCA	S000403
TATCCAYMOTIFOSRAMY3D	1221 (-) TATCCAY	S000256
TBOXATGAPB	302 (+) ACTTTG	S000383
TBOXATGAPB	780 (+) ACTTTG	S000383
TBOXATGAPB	917 (+) ACTTTG	S000383
TELOBOXATEEF1AA1	649 (+) AAACCCTAA	S000308
UP2ATMSD	649 (+) AAACCCTA	S000472
WBBOXPCWRKY1	678 (-) TTTGACY	S000310
WBOXATNPR1	265 (+) TTGAC	S000390
WBOXATNPR1	1186 (+) TTGAC	S000390
WBOXATNPR1	93 (-) TTGAC	S000390
WBOXATNPR1	679 (-) TTGAC	S000390
WBOXHVISO1	92 (-) TGACT	S000442
WBOXHVISO1	572 (-) TGACT	S000442
WBOXNTCHN48	572 (-) CTGACY	S000508
WBOXNTERF3	92 (-) TGACY	S000457
WBOXNTERF3	572 (-) TGACY	S000457
WBOXNTERF3	678 (-) TGACY	S000457
WBOXNTERF3	1414 (-) TGACY	S000457
WRKY71OS	56 (+) TGAC	S000447
WRKY71OS	266 (+) TGAC	S000447
WRKY71OS	1187 (+) TGAC	S000447
WRKY71OS	93 (-) TGAC	S000447
WRKY71OS	573 (-) TGAC	S000447
WRKY71OS	679 (-) TGAC	S000447
WRKY71OS	1415 (-) TGAC	S000447
WRKY71OS	1446 (-) TGAC	S000447

Segment of chromosome 8 upstream of LJFgene8 start codon:

ATTATTCGAATCAATTTGGGTCAATTCGAATTTGTGGATGACTTATACCC
ATAAACAACCCTACATAGAAAATTATTGATGCAAGCATTAGTTGTTAATT
ATTGAAGAGAGTCTTCTCCCGACCAAGAGCTTCTAGTCCTTTCAACATTC
TTTCGTGCAAGTAGAAAGAAAATACTAGTTTCACCTATCTAGTCTTAAAT
TCGTTGAACTGTGTATTTGTAATTTTAGCAAGGTTAGGGTTTAGGGAGGA
CTACTAGATAGGTGACCTAGCCTTGCTCTATAATATATTATGCTAAAAGG
AATTATGATTGAAAATATACCACTATATTATGGATGGATCCAAATGGGAG
AAGGATTAATTCCGAAAAATAAAAGCTCCAAATATGTTCTGGCTACACAA
AAAAATCATTAAAAAAAACCTTATTGGTTGATACTTTTAAAAAAATATAG
TTTTTTTAAAGATTTATTTCACATAAATTTTATCTTTTGAAGATTTATTT
TATATTAAACAAGCCTTTCAAAAAAATTTAATTTAAAAAATATTTCATAT
TACAAAACTTTAAAATTTTAATTTATTTTTTCATATTGTTTTTATTCTAT

AATACTATTTTATTTATTATTAATTTTTAGGTTATGTATATAATTTAATC
TCATATATACATTATCAATTGCATGTGTTCCCGCCGTTTACACATTCTTA
ATTTAATCATTATATGTAAAATATATTCTTAATTAATTTGATCTTTATAA
TAATTAAGAAATGCTCTACAAGTATATCACCAAATTAAGAATGAATTACA
TATCTAGAATTCAAATTAAGAATGTGTTGCACATACAAAGATTAAATTAA
GAATGCATAAAAGACATCATAAAGATAAATCAAATCGATAGTATTTGTAC
ATACAAAGACAAATTTAATGAGTTTTTACTACAAGGACTAGATTAACTAC
ACATTTAACTTACTTTTTATTTGTTATAATCTATTTGTATGATTTTGATA
ACTTCTTTTCATAATTTTTCAATGATTATGATGATATTTCACTATTTTTT
TTATCTTTCATCTTTTTAGTATCTTTACAATCTATTTTTATATTAATATT
TGTAATTTATTGCAAAAAATTTAGTAATTATAATATGGATTTTCTATGAA
TATACTATTGCATACTATAACATTAGTATTATTTACTAATTACAAATAAA
ATGCATATAAATAAAGAGTTGGAAATATTATTTCTAAAACAGCATGATAT
ACCTCTAGTGTCCACTTAATGGTGAGATTTGAAATAATATGTATGAGATT
CTAAGTTCAAATATTAAGTGTCATTGTTAAAGAAAAAACAGTATGATATG
CTGGAATTTACTAACTATAATCTTTATAATAAGTTTAGCATTTAGTGTAT
ATTGACTTGAAAGATTCTTGTAGCAATTTGCAGCAGTTTTATATAGATAG
TAACATTCTTAAATTACGTTTATAAGTTCCTTCATTTTCAGCTCCGAACA
ATGGCTTCTTCGTTCTCCTTCTGCACCCTCAAGTTTCGCACCAAACCCAA

-10PEHVPSBD	593 (+) TATTCT		S000392
-10PEHVPSBD	724 (+) TATTCT		S000392
-300CORE	1073 (-) TGTAAAG		S000001
-300ELEMENT	715 (+) TGHAAARK		S000122
-300ELEMENT	1014 (-) TGHAAARK		S000122
ACGTATERD1	1466 (+) ACGT		S000415
ACGTATERD1	1466 (-) ACGT		S000415
AMYBOX1	970 (-) TAACARA		S000020
ARE1	261 (+) RGTGACNNNGC		S000022
ARR1AT	460 (+) NGATT		S000454
ARR1AT	491 (+) NGATT		S000454
ARR1AT	839 (+) NGATT		S000454
ARR1AT	940 (+) NGATT		S000454
ARR1AT	1275 (+) NGATT		S000454
ARR1AT	1296 (+) NGATT		S000454
ARR1AT	1412 (+) NGATT		S000454
ARR1AT	353 (+) NGATT		S000454
ARR1AT	1137 (+) NGATT		S000454
ARR1AT	306 (+) NGATT		S000454
ARR1AT	990 (+) NGATT		S000454
ARR1AT	1023 (+) NGATT		S000454
ARR1AT	9 (-) NGATT		S000454
ARR1AT	404 (-) NGATT		S000454
ARR1AT	647 (-) NGATT		S000454
ARR1AT	705 (-) NGATT		S000454
ARR1AT	878 (-) NGATT		S000454
ARR1AT	883 (-) NGATT		S000454
ARR1AT	978 (-) NGATT		S000454
ARR1AT	1079 (-) NGATT		S000454
ARR1AT	1369 (-) NGATT		S000454
BIHD1OS	1319 (+) TGTCA		S000498
BOXIINTPATPB	65 (+) ATAGAA		S000296
BOXIINTPATPB	595 (-) ATAGAA		S000296
BOXIINTPATPB	1142 (-) ATAGAA		S000296
CAATBOX1	12 (+) CAAT		S000028

CAATBOX1	22 (+) CAAT	S000028
CAATBOX1	666 (+) CAAT	S000028
CAATBOX1	1020 (+) CAAT	S000028
CAATBOX1	1078 (+) CAAT	S000028
CAATBOX1	1424 (+) CAAT	S000028
CAATBOX1	1499 (+) CAAT	S000028
CAATBOX1	75 (-) CAAT	S000028
CAATBOX1	101 (-) CAAT	S000028
CAATBOX1	308 (-) CAAT	S000028
CAATBOX1	423 (-) CAAT	S000028
CAATBOX1	585 (-) CAAT	S000028
CAATBOX1	668 (-) CAAT	S000028
CAATBOX1	1109 (-) CAAT	S000028
CAATBOX1	1157 (-) CAAT	S000028
CAATBOX1	1323 (-) CAAT	S000028
CAATBOX1	1401 (-) CAAT	S000028
CACTFTPPCA1	321 (+) YACT	S000449
CACTFTPPCA1	1040 (+) YACT	S000449
CACTFTPPCA1	1263 (+) YACT	S000449
CACTFTPPCA1	173 (+) YACT	S000449
CACTFTPPCA1	252 (+) YACT	S000449
CACTFTPPCA1	432 (+) YACT	S000449
CACTFTPPCA1	603 (+) YACT	S000449
CACTFTPPCA1	927 (+) YACT	S000449
CACTFTPPCA1	961 (+) YACT	S000449
CACTFTPPCA1	1153 (+) YACT	S000449
CACTFTPPCA1	1163 (+) YACT	S000449
CACTFTPPCA1	1184 (+) YACT	S000449
CACTFTPPCA1	1359 (+) YACT	S000449
CACTFTPPCA1	160 (-) YACT	S000449
CACTFTPPCA1	771 (-) YACT	S000449
CACTFTPPCA1	890 (-) YACT	S000449
CACTFTPPCA1	1068 (-) YACT	S000449
CACTFTPPCA1	1123 (-) YACT	S000449
CACTFTPPCA1	1175 (-) YACT	S000449
CACTFTPPCA1	1257 (-) YACT	S000449
CACTFTPPCA1	1317 (-) YACT	S000449
CACTFTPPCA1	1340 (-) YACT	S000449
CACTFTPPCA1	1394 (-) YACT	S000449
CACTFTPPCA1	1449 (-) YACT	S000449
CAREOSREP1	1216 (-) CAACTC	S000421
CARGCW8GAT	323 (+) CWWWWWWWWG	S000431
CARGCW8GAT	323 (-) CWWWWWWWWG	S000431
CCAATBOX1	423 (-) CCAAT	S000030
CIACADIANLELHC	398 (+) CAANNNNATC	S000252
CIACADIANLELHC	769 (+) CAANNNNATC	S000252
CURECORECR	897 (+) GTAC	S000493
CURECORECR	897 (-) GTAC	S000493
DOFCOREZM	165 (+) AAAG	S000265
DOFCOREZM	296 (+) AAAG	S000265
DOFCOREZM	372 (+) AAAG	S000265
DOFCOREZM	458 (+) AAAG	S000265
DOFCOREZM	837 (+) AAAG	S000265
DOFCOREZM	860 (+) AAAG	S000265

DOFCOREZM	871 (+) AAAG	S000265
DOFCOREZM	905 (+) AAAG	S000265
DOFCOREZM	1213 (+) AAAG	S000265
DOFCOREZM	1329 (+) AAAG	S000265
DOFCOREZM	1410 (+) AAAG	S000265
DOFCOREZM	139 (-) AAAG	S000265
DOFCOREZM	150 (-) AAAG	S000265
DOFCOREZM	434 (-) AAAG	S000265
DOFCOREZM	484 (-) AAAG	S000265
DOFCOREZM	515 (-) AAAG	S000265
DOFCOREZM	558 (-) AAAG	S000265
DOFCOREZM	743 (-) AAAG	S000265
DOFCOREZM	963 (-) AAAG	S000265
DOFCOREZM	1005 (-) AAAG	S000265
DOFCOREZM	1055 (-) AAAG	S000265
DOFCOREZM	1062 (-) AAAG	S000265
DOFCOREZM	1073 (-) AAAG	S000265
DOFCOREZM	1372 (-) AAAG	S000265
DPBFCOREDCDC3	672 (-) ACACNNG	S000292
DPBFCOREDCDC3	1255 (-) ACACNNG	S000292
EBOXBNNAPA	341 (+) CANNTG	S000144
EBOXBNNAPA	666 (+) CANNTG	S000144
EBOXBNNAPA	672 (+) CANNTG	S000144
EBOXBNNAPA	341 (-) CANNTG	S000144
EBOXBNNAPA	666 (-) CANNTG	S000144
EBOXBNNAPA	672 (-) CANNTG	S000144
EECCRCAH1	793 (+) GANTTNC	S000494
EECCRCAH1	1138 (+) GANTTNC	S000494
ELRECOREPCRP1	19 (-) TTGACC	S000142
ERELEE4	808 (+) AWTTCAAA	S000037
ERELEE4	1278 (-) AWTTCAAA	S000037
GAREAT	970 (-) TAACAAR	S000439
GATABOX	257 (+) GATA	S000039
GATABOX	430 (+) GATA	S000039
GATABOX	874 (+) GATA	S000039
GATABOX	887 (+) GATA	S000039
GATABOX	997 (+) GATA	S000039
GATABOX	1033 (+) GATA	S000039
GATABOX	1246 (+) GATA	S000039
GATABOX	1345 (+) GATA	S000039
GATABOX	1446 (+) GATA	S000039
GATABOX	186 (-) GATA	S000039
GATABOX	481 (-) GATA	S000039
GATABOX	663 (-) GATA	S000039
GATABOX	775 (-) GATA	S000039
GATABOX	801 (-) GATA	S000039
GATABOX	1052 (-) GATA	S000039
GATABOX	1070 (-) GATA	S000039
GT1CONSENSUS	68 (+) GRWAAW	S000198
GT1CONSENSUS	168 (+) GRWAAW	S000198
GT1CONSENSUS	311 (+) GRWAAW	S000198
GT1CONSENSUS	364 (+) GRWAAW	S000198
GT1CONSENSUS	874 (+) GRWAAW	S000198
GT1CONSENSUS	1221 (+) GRWAAW	S000198

GT1CONSENSUS	1332 (+) GRWAAW	S000198
GT1CONSENSUS	661 (-) GRWAAW	S000198
GT1CONSENSUS	1139 (-) GRWAAW	S000198
GT1CONSENSUS	1484 (-) GRWAAW	S000198
GT1CONSENSUS	479 (-) GRWAAW	S000198
GT1CONSENSUS	577 (-) GRWAAW	S000198
GT1CONSENSUS	1015 (-) GRWAAW	S000198
GT1CONSENSUS	1050 (-) GRWAAW	S000198
GT1GMSCAM4	364 (+) GAAAAA	S000453
GT1GMSCAM4	1332 (+) GAAAAA	S000453
GT1GMSCAM4	577 (-) GAAAAA	S000453
GT1GMSCAM4	1015 (-) GAAAAA	S000453
GTGANTG10	262 (+) GTGA	S000378
GTGANTG10	1272 (+) GTGA	S000378
GTGANTG10	181 (-) GTGA	S000378
GTGANTG10	469 (-) GTGA	S000378
GTGANTG10	777 (-) GTGA	S000378
GTGANTG10	1039 (-) GTGA	S000378
HDZIP2ATATHB2	704 (+) TAATMATTA	S000373
HDZIP2ATATHB2	748 (+) TAATMATTA	S000373
HSELIKENTACIDICPR1	1406 (+) CNNGAANNNTTCNNG	S000056
HSELIKENTACIDICPR1	1406 (-) CNNGAANNNTTCNNG	S000056
IBOXCORE	874 (+) GATAA	S000199
IBOXCORE	997 (+) GATAA	S000199
IBOXCORE	480 (-) GATAA	S000199
IBOXCORE	662 (-) GATAA	S000199
IBOXCORE	1051 (-) GATAA	S000199
INRNTPSADB	1481 (+) YTCANTYY	S000395
INRNTPSADB	788 (-) YTCANTYY	S000395
LECPLEACS2	717 (+) TAAAATAT	S000465
LTRE1HVBLT49	362 (+) CCGAAA	S000250
LTRECOREATCOR15	119 (+) CCGAC	S000153
MARABOX1	609 (-) AATAAAYAAA	S000063
MARTBOX	365 (-) TTWTWTTWTT	S000067
MYBCOREATCYCB1	684 (-) AACGG	S000502
MYBGAHV	970 (-) TAACAAA	S000181
MYBPLANT	1539 (+) MACCWAMC	S000167
MYCATERD1	672 (+) CATGTG	S000413
MYCATRD22	672 (-) CACATG	S000174
MYCCONSENSUSAT	341 (+) CANNTG	S000407
MYCCONSENSUSAT	666 (+) CANNTG	S000407
MYCCONSENSUSAT	672 (+) CANNTG	S000407
MYCCONSENSUSAT	341 (-) CANNTG	S000407
MYCCONSENSUSAT	666 (-) CANNTG	S000407
MYCCONSENSUSAT	672 (-) CANNTG	S000407
NAPINMOTIFBN	689 (+) TACACAT	S000070
NAPINMOTIFBN	948 (+) TACACAT	S000070
NODCON1GM	458 (+) AAAGAT	S000461
NODCON1GM	837 (+) AAAGAT	S000461
NODCON1GM	871 (+) AAAGAT	S000461
NODCON1GM	1410 (+) AAAGAT	S000461
NODCON1GM	482 (-) AAAGAT	S000461

NODCON1GM	741 (-) AAAGAT	S000461
NODCON1GM	1053 (-) AAAGAT	S000461
NODCON1GM	1060 (-) AAAGAT	S000461
NODCON1GM	1071 (-) AAAGAT	S000461
NODCON1GM	1370 (-) AAAGAT	S000461
NODCON2GM	105 (-) CTCTT	S000462
NODCON2GM	125 (-) CTCTT	S000462
NODCON2GM	1214 (-) CTCTT	S000462
NTBBF1ARROLB	557 (+) ACTTTA	S000273
OSE1ROOTNODULE	458 (+) AAAGAT	S000467
OSE1ROOTNODULE	837 (+) AAAGAT	S000467
OSE1ROOTNODULE	871 (+) AAAGAT	S000467
OSE1ROOTNODULE	1410 (+) AAAGAT	S000467
OSE1ROOTNODULE	482 (-) AAAGAT	S000467
OSE1ROOTNODULE	741 (-) AAAGAT	S000467
OSE1ROOTNODULE	1053 (-) AAAGAT	S000467
OSE1ROOTNODULE	1060 (-) AAAGAT	S000467
OSE1ROOTNODULE	1071 (-) AAAGAT	S000467
OSE1ROOTNODULE	1370 (-) AAAGAT	S000467
OSE2ROOTNODULE	105 (-) CTCTT	S000468
OSE2ROOTNODULE	125 (-) CTCTT	S000468
OSE2ROOTNODULE	1214 (-) CTCTT	S000468
P1BS	1148 (+) GNATATNC	S000459
P1BS	1148 (-) GNATATNC	S000459
POLASIG1	368 (+) AATAAA	S000080
POLASIG1	1195 (+) AATAAA	S000080
POLASIG1	1210 (+) AATAAA	S000080
POLASIG1	463 (-) AATAAA	S000080
POLASIG1	494 (-) AATAAA	S000080
POLASIG1	572 (-) AATAAA	S000080
POLASIG1	591 (-) AATAAA	S000080
POLASIG1	609 (-) AATAAA	S000080
POLASIG1	613 (-) AATAAA	S000080
POLASIG1	966 (-) AATAAA	S000080
POLASIG1	1106 (-) AATAAA	S000080
POLASIG2	527 (-) AATTAAA	S000081
POLASIG2	567 (-) AATTAAA	S000081
POLASIG3	749 (+) AATAAT	S000088
POLASIG3	1283 (+) AATAAT	S000088
POLASIG3	1 (-) AATAAT	S000088
POLASIG3	72 (-) AATAAT	S000088
POLASIG3	98 (-) AATAAT	S000088
POLASIG3	616 (-) AATAAT	S000088
POLASIG3	1178 (-) AATAAT	S000088
POLASIG3	1227 (-) AATAAT	S000088
POLLEN1LELAT52	67 (+) AGAAA	S000245
POLLEN1LELAT52	163 (+) AGAAA	S000245
POLLEN1LELAT52	167 (+) AGAAA	S000245
POLLEN1LELAT52	757 (+) AGAAA	S000245
POLLEN1LELAT52	1331 (+) AGAAA	S000245
POLLEN1LELAT52	1141 (-) AGAAA	S000245
POLLEN1LELAT52	1231 (-) AGAAA	S000245
PREATPRODH	918 (-) ACTCAT	S000450
PRECONSCRHSP70A	119 (+)	S000506

	SCGAYNRNNNNNNNNNNNNNNHD	
PYRIMIDINEBOXOSRAMY1A	295 (-) CCTTTT	S000259
QELEMENTZMZM13	263 (-) AGGTCA	S000254
RAV1AAT	143 (+) CAACA	S000314
RAV1AAT	825 (-) CAACA	S000314
REALPHALGLHCB21	424 (-) AACCAA	S000362
RHERPATEXPA7	153 (-) KCACGW	S000512
ROOTMOTIFTAPOX1	285 (+) ATATT	S000098
ROOTMOTIFTAPOX1	325 (+) ATATT	S000098
ROOTMOTIFTAPOX1	502 (+) ATATT	S000098
ROOTMOTIFTAPOX1	540 (+) ATATT	S000098
ROOTMOTIFTAPOX1	547 (+) ATATT	S000098
ROOTMOTIFTAPOX1	583 (+) ATATT	S000098
ROOTMOTIFTAPOX1	723 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1034 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1090 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1096 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1225 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1311 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1399 (+) ATATT	S000098
ROOTMOTIFTAPOX1	282 (-) ATATT	S000098
ROOTMOTIFTAPOX1	314 (-) ATATT	S000098
ROOTMOTIFTAPOX1	381 (-) ATATT	S000098
ROOTMOTIFTAPOX1	444 (-) ATATT	S000098
ROOTMOTIFTAPOX1	539 (-) ATATT	S000098
ROOTMOTIFTAPOX1	720 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1095 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1132 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1149 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1224 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1286 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1310 (-) ATATT	S000098
RYREPEATBNNAPA	670 (-) CATGCA	S000264
SEF3MOTIFGM	1544 (+) AACCCA	S000115
SEF4MOTIFGM7S	623 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	1084 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	588 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	922 (+) RTTTTTR	S000103
SORLREP3AT	653 (-) TGTATATAT	S000488
SP8BFIBSP8AIB	208 (+) ACTGTGTA	S000183
SP8BFIBSP8BIB	603 (+) TACTATT	S000184
SP8BFIBSP8BIB	1153 (+) TACTATT	S000184
SURE2STPAT21	1172 (-) AATACTAAT	S000185
TAAAGSTKST1	457 (+) TAAAG	S000387
TAAAGSTKST1	870 (+) TAAAG	S000387
TAAAGSTKST1	1212 (+) TAAAG	S000387
TAAAGSTKST1	1328 (+) TAAAG	S000387
TAAAGSTKST1	558 (-) TAAAG	S000387
TAAAGSTKST1	743 (-) TAAAG	S000387
TAAAGSTKST1	1073 (-) TAAAG	S000387
TAAAGSTKST1	1372 (-) TAAAG	S000387
TATABOX2	1206 (+) TATAAAT	S000109
TATABOX3	618 (+) TATTAAT	S000110

TATABOX3	1091 (+) TATTAAT	S000110
TATABOX3	1092 (-) TATTAAT	S000110
TATABOX4	637 (+) TATATAA	S000111
TATABOX4	1439 (-) TATATAA	S000111
TATABOX5	464 (+) TTATTT	S000203
TATABOX5	495 (+) TTATTT	S000203
TATABOX5	573 (+) TTATTT	S000203
TATABOX5	610 (+) TTATTT	S000203
TATABOX5	967 (+) TTATTT	S000203
TATABOX5	1179 (+) TTATTT	S000203
TATABOX5	1228 (+) TTATTT	S000203
TATABOX5	367 (-) TTATTT	S000203
TATABOX5	1194 (-) TTATTT	S000203
TATABOX5	1209 (-) TTATTT	S000203
TATABOX5	1282 (-) TTATTT	S000203
TATAPVTRNALEU	1438 (+) TTTATATA	S000340
TELOBOXATEEF1AA1	234 (-) AAACCCTAA	S000308
UP2ATMSD	235 (-) AAACCCTA	S000472
WBOXATNPR1	1402 (+) TTGAC	S000390
WBOXATNPR1	20 (-) TTGAC	S000390
WBOXHVISO1	39 (+) TGACT	S000442
WBOXHVISO1	1403 (+) TGACT	S000442
WBOXNTERF3	39 (+) TGACY	S000457
WBOXNTERF3	263 (+) TGACY	S000457
WBOXNTERF3	1403 (+) TGACY	S000457
WBOXNTERF3	19 (-) TGACY	S000457
WRKY71OS	39 (+) TGAC	S000447
WRKY71OS	263 (+) TGAC	S000447
WRKY71OS	1403 (+) TGAC	S000447
WRKY71OS	20 (-) TGAC	S000447
WRKY71OS	1320 (-) TGAC	S000447
WUSATAg	1266 (+) TTAATGG	S000433

Segment of chromosome 9 upstream of LJFgene9 start codon:

AGAAGAAAAGTTGAATAGTATATATATGAACATTGAAATTATTTTTTAAG
AAATTCACAAAGCTTTGATAGAAAATAAGTTCATGAATTGTTTTCGTTCA
AAGAAAATGTGAGACATAAAAGAAACAAAAGCACAAAACACAATTTTTTT
TTTATAGTATTTCACTTAACTTAAAAGTTACATTTAGTTCTCTTTCAATC
TTAAAAAGATTTCACTAACCAAATTTAAAAAAAGTATTATTCATACCACT
TATGGTTATAATAAATGTTTTTTATACTTCTAACTACAACAAACAATCTC
AATCTGAGATAAAAAAATTCAAAGACTCCCTTCAAAACTAAATATAAATA
AAATCAATTAACCCTTTATATGTTTAATTAAGAAAATTAGCTAATATCAT
TTAATTATATTATTTGCATTTAAGAAATATTTTTTTTTCTAAAGTATATT
TCATTCAAATTTGATGTCAAAAATGAAAGAAATCACTGGAAATAGAATAT
CATTAATTGAAGGATATATTTTTGAGATATCATCATTAAATAATTAAAAT
TTTATTTATATTTAAGTTCAATTTTTTGTTTGTTAATTATACTTAGATAC
AAATGAAGTATTCTTCAACATAAAGTCACTTCTATCCTTCACAAGCAACA
AGCAACAAAGTATTTGAAAAATTACAAGAATCTAATGAGGAAGTGAGGAT
ATGTACGTTAAGCATAGATACAACAGGGCAGGGAGAGTTTATGCTCAAGG
TGTTTCATTCAAGAAAAGTATTAGAGGAGTGTGAAATGGAGCCTTTGTGA
TTTCCCCATTCCCATCTTCCTCTGTAATCTTAATTTATACTTTTAATTTT
TAATACAACGATATTTTATTATTAAAATATTATTTTGTTAATCGATAGAC
ACAGTTGTCTAAAATTTTATTATTGCTTTCATACACTGAGGGAACAATTT

ATTTAAGAAAAGAATTTGTATTTGATTTTTCTATATATAAATTTTTCTTA
CATCAACCATAATCACTCATTGTAATTAAAAATTAAAATTTCAATAAGTT
AATCAACGCTCACGTTTTCTAACACAAATAATAAAAGAGAATATTTTTTT
GGTATATGTGTTTTAATTATAATAACTAATAAATCCACAACGTGTATGCC
ACTTCCCATTGTCCCGCACATACACTTGAAAAAAGTCCAATTTGCATTTT
AGCATTGGTTCGCACCTAAGGCACCTTCCCAATTCAGCTTCTAACGATGA
CACTTTGTAGCACGTTTTCCAACTTCAACATTCACATATTAAAAAACAAC
AAGGGTTCCTTTTCTCGTCGATTTCAACTCTCTCAGAAGCTGGATGACGA
TAATTTCATTGATAAAATCAAACGAAGGTTCTCACTGATTCTCCCTTTAA
TTTGCCACCTCACATGAATTGTATAATATATATTTATATTTATGCTTGAC
CTTGAATTGTTCCTATCTTAAAGAGAGCTCATACTGGAAAGTGGAGAATT
AGCAACCATTGGTGCCATCTTCAACTTTAGGTACACTGCTTTATTGTTTT
CACAATGAGAATAGGTGACCAAAATTTAATTAATTCTTATAATATTTGA

-10PEHVPSBD	609	(+) TATTCT	S000392
-10PEHVPSBD	494	(-) TATTCT	S000392
-10PEHVPSBD	1088	(-) TATTCT	S000392
-10PEHVPSBD	1558	(-) TATTCT	S000392
-300ELEMENT	665	(+) TGHAAARK	S000122
AACACOREOSGLUB1	288	(+) AACAAAC	S000353
AACACOREOSGLUB1	578	(-) AACAAAC	S000353
ABRELATERD1	1140	(+) ACGTG	S000414
ABRELATERD1	1061	(-) ACGTG	S000414
ABRELATERD1	1261	(-) ACGTG	S000414
ABRERATCAL	1139	(+) MACGYGB	S000507
ABRERATCAL	1260	(-) MACGYGB	S000507
ACGTATERD1	705	(+) ACGT	S000415
ACGTATERD1	1062	(+) ACGT	S000415
ACGTATERD1	1140	(+) ACGT	S000415
ACGTATERD1	1262	(+) ACGT	S000415
ACGTATERD1	705	(-) ACGT	S000415
ACGTATERD1	1062	(-) ACGT	S000415
ACGTATERD1	1140	(-) ACGT	S000415
ACGTATERD1	1262	(-) ACGT	S000415
AMYBOX1	579	(-) TAACARA	S000020
AMYBOX1	884	(-) TAACARA	S000020
ANAERO1CONSENSUS	123	(+) AAACAAA	S000477
ANAERO1CONSENSS	575	(-) AAACAAA	S000477
ARFAT	111	(-) TGTCTC	S000270
ARR1AT	207	(+) NGATT	S000454
ARR1AT	1319	(+) NGATT	S000454
ARR1AT	798	(+) NGATT	S000454
ARR1AT	973	(+) NGATT	S000454
ARR1AT	1386	(+) NGATT	S000454
ARR1AT	197	(-) NGATT	S000454
ARR1AT	295	(-) NGATT	S000454
ARR1AT	301	(-) NGATT	S000454
ARR1AT	352	(-) NGATT	S000454
ARR1AT	481	(-) NGATT	S000454
ARR1AT	679	(-) NGATT	S000454
ARR1AT	826	(-) NGATT	S000454
ARR1AT	890	(-) NGATT	S000454
ARR1AT	1011	(-) NGATT	S000454
ARR1AT	1051	(-) NGATT	S000454
ARR1AT	1132	(-) NGATT	S000045

ARR1AT	1366 (-) NGATT	S000454
ASF1MOTIFCAMV	1345 (+) TGACG	S000024
BIHD1OS	465 (+) TGTCA	S000498
BIHD1OS	1248 (-) TGTCA	S000498
BOXIINTPATPB	68 (+) ATAGAA	S000296
BOXIINTPATPB	492 (+) ATAGAA	S000296
BOXIINTPATPB	630 (-) ATAGAA	S000296
BOXIINTPATPB	979 (-) ATAGAA	S000296
BOXLCOREDCPAL	1223 (+) ACCWWCC	S000492
BP5OSWX	1138 (+) CAACGTG	S000436
CAATBOX1	141 (+) CAAT	S000028
CAATBOX1	196 (+) CAAT	S000028
CAATBOX1	294 (+) CAAT	S000028
CAATBOX1	300 (+) CAAT	S000028
CAATBOX1	355 (+) CAAT	S000028
CAATBOX1	569 (+) CAAT	S000028
CAATBOX1	945 (+) CAAT	S000028
CAATBOX1	1042 (+) CAAT	S000028
CAATBOX1	1188 (+) CAAT	S000028
CAATBOX1	1230 (+) CAAT	S000028
CAATBOX1	1553 (+) CAAT	S000028
CAATBOX1	32 (-) CAAT	S000028
CAATBOX1	87 (-) CAAT	S000028
CAATBOX1	506 (-) CAAT	S000028
CAATBOX1	922 (-) CAAT	S000028
CAATBOX1	1019 (-) CAAT	S000028
CAATBOX1	1158 (-) CAAT	S000028
CAATBOX1	1204 (-) CAAT	S000028
CAATBOX1	1358 (-) CAAT	S000028
CAATBOX1	1418 (-) CAAT	S000028
CAATBOX1	1456 (-) CAAT	S000028
CAATBOX1	1508 (-) CAAT	S000028
CAATBOX1	1543 (-) CAAT	S000028
CACTFTPPCA1	163 (+) YACT	S000449
CACTFTPPCA1	213 (+) YACT	S000449
CACTFTPPCA1	247 (+) YACT	S000449
CACTFTPPCA1	484 (+) YACT	S000449
CACTFTPPCA1	627 (+) YACT	S000449
CACTFTPPCA1	934 (+) YACT	S000449
CACTFTPPCA1	1014 (+) YACT	S000449
CACTFTPPCA1	1150 (+) YACT	S000449
CACTFTPPCA1	1173 (+) YACT	S000449
CACTFTPPCA1	1251 (+) YACT	S000449
CACTFTPPCA1	1383 (+) YACT	S000449
CACTFTPPCA1	1534 (+) YACT	S000449
CACTFTPPCA1	275 (+) YACT	S000449
CACTFTPPCA1	590 (+) YACT	S000449
CACTFTPPCA1	838 (+) YACT	S000449
CACTFTPPCA1	1482 (+) YACT	S000449
CACTFTPPCA1	17 (-) YACT	S000449
CACTFTPPCA1	156 (-) YACT	S000449
CACTFTPPCA1	233 (-) YACT	S000449
CACTFTPPCA1	443 (-) YACT	S000449
CACTFTPPCA1	607 (-) YACT	S000449

CACTFTPPCA1	659 (-) YACT	S000449
CACTFTPPCA1	692 (-) YACT	S000449
CACTFTPPCA1	767 (-) YACT	S000449
CACTFTPPCA1	778 (-) YACT	S000449
CACTFTPPCA1	1490 (-) YACT	S000449
CANBNNAPA	1069 (+) CNAACAC	S000148
CAREOSREP1	1325 (+) CAACTC	S000421
CARGCW8GAT	569 (+) CWWWWWWWWG	S000431
CARGCW8GAT	1395 (+) CWWWWWWWWG	S000431
CARGCW8GAT	569 (-) CWWWWWWWWG	S000431
CARGCW8GAT	1395 (-) CWWWWWWWWG	S000431
CBFHV	1317 (-) RYCGAC	S000497
CCA1ATLHCB1	292 (+) AAMAATCT	S000149
CCAATBOX1	1187 (+) CCAAT	S000030
CCAATBOX1	1229 (+) CCAAT	S000030
CCAATBOX1	1204 (-) CCAAT	S000030
CCAATBOX1	1508 (-) CCAAT	S000030
CGACGOSAMY3	1316 (-) CGACG	S000205
CPBCSPOR	769 (+) TATTAG	S000491
CPBCSPOR	391 (-) TATTAG	S000491
CPBCSPOR	1126 (-) TATTAG	S000491
CURECORECR	703 (+) GTAC	S000493
CURECORECR	1531 (+) GTAC	S000493
CURECORECR	703 (-) GTAC	S000493
CURECORECR	1531 (-) GTAC	S000493
DOFCOREZM	7 (+) AAAG	S000265
DOFCOREZM	59 (+) AAAG	S000265
DOFCOREZM	100 (+) AAAG	S000265
DOFCOREZM	119 (+) AAAG	S000265
DOFCOREZM	128 (+) AAAG	S000265
DOFCOREZM	174 (+) AAAG	S000265
DOFCOREZM	205 (+) AAAG	S000265
DOFCOREZM	231 (+) AAAG	S000265
DOFCOREZM	321 (+) AAAG	S000265
DOFCOREZM	441 (+) AAAG	S000265
DOFCOREZM	476 (+) AAAG	S000265
DOFCOREZM	622 (+) AAAG	S000265
DOFCOREZM	657 (+) AAAG	S000265
DOFCOREZM	765 (+) AAAG	S000265
DOFCOREZM	959 (+) AAAG	S000265
DOFCOREZM	1084 (+) AAAG	S000265
DOFCOREZM	1182 (+) AAAG	S000265
DOFCOREZM	1470 (+) AAAG	S000265
DOFCOREZM	1488 (+) AAAG	S000265
DOFCOREZM	63 (-) AAAG	S000265
DOFCOREZM	192 (-) AAAG	S000265
DOFCOREZM	364 (-) AAAG	S000265
DOFCOREZM	793 (-) AAAG	S000265
DOFCOREZM	840 (-) AAAG	S000265
DOFCOREZM	926 (-) AAAG	S000265
DOFCOREZM	1253 (-) AAAG	S000265
DOFCOREZM	1309 (-) AAAG	S000265
DOFCOREZM	1395 (-) AAAG	S000265
DOFCOREZM	1525 (-) AAAG	S000265

DOFCOREZM	1539 (-) AAAG	S000265
DPBFCOREDCDC3	1172 (+) ACACNNG	S000292
EBOXBNNAPA	600 (+) CANNTG	S000144
EBOXBNNAPA	902 (+) CANNTG	S000144
EBOXBNNAPA	1173 (+) CANNTG	S000144
EBOXBNNAPA	1411 (+) CANNTG	S000144
EBOXBNNAPA	600 (-) CANNTG	S000144
EBOXBNNAPA	902 (-) CANNTG	S000144
EBOXBNNAPA	1173 (-) CANNTG	S000144
EBOXBNNAPA	1411 (-) CANNTG	S000144
EECCRCAH1	799 (+) GANTTNC	S000494
EECCRCAH1	50 (-) GANTTNC	S000494
ELRECOREPCRP1	1446 (+) TTGACC	S000142
ERELEE4	316 (+) AWTTCAAA	S000037
GARE2OSREP1	704 (-) TAACGTA	S000420
GAREAT	579 (-) TAACAAR	S000439
GAREAT	884 (-) TAACAAR	S000439
GATABOX	67 (+) GATA	S000039
GATABOX	308 (+) GATA	S000039
GATABOX	513 (+) GATA	S000039
GATABOX	526 (+) GATA	S000039
GATABOX	596 (+) GATA	S000039
GATABOX	698 (+) GATA	S000039
GATABOX	717 (+) GATA	S000039
GATABOX	860 (+) GATA	S000039
GATABOX	894 (+) GATA	S000039
GATABOX	1349 (+) GATA	S000039
GATABOX	1361 (+) GATA	S000039
GATABOX	395 (-) GATA	S000039
GATABOX	498 (-) GATA	S000039
GATABOX	528 (-) GATA	S000039
GATABOX	633 (-) GATA	S000039
GATABOX	1464 (-) GATA	S000039
GT1CONSENSUS	71 (+) GRWAAW	S000198
GT1CONSENSUS	103 (+) GRWAAW	S000198
GT1CONSENSUS	308 (+) GRWAAW	S000198
GT1CONSENSUS	382 (+) GRWAAW	S000198
GT1CONSENSUS	488 (+) GRWAAW	S000198
GT1CONSENSUS	666 (+) GRWAAW	S000198
GT1CONSENSUS	1178 (+) GRWAAW	S000198
GT1CONSENSUS	1349 (+) GRWAAW	S000198
GT1CONSENSUS	1361 (+) GRWAAW	S000198
GT1CONSENSUS	800 (-) GRWAAW	S000198
GT1CONSENSUS	434 (-) GRWAAW	S000198
GT1CONSENSUS	976 (-) GRWAAW	S000198
GT1CONSENSUS	992 (-) GRWAAW	S000198
GT1CONSENSUS	1265 (-) GRWAAW	S000198
GT1CORE	358 (-) GGTTAA	S000125
GT1GMSCAM4	666 (+) GAAAAA	S000453
GT1GMSCAM4	1178 (+) GAAAAA	S000453
GT1GMSCAM4	434 (-) GAAAAA	S000453
GT1GMSCAM4	976 (-) GAAAAA	S000453
GT1GMSCA4	992 (-) GAAAAA	S000453
GTGANTG10	109 (+) GTGA	S000378

GTGANTG10	693 (+) GTGA	S000378
GTGANTG10	781 (+) GTGA	S000378
GTGANTG10	797 (+) GTGA	S000378
GTGANTG10	1565 (+) GTGA	S000378
GTGANTG10	55 (-) GTGA	S000378
GTGANTG10	162 (-) GTGA	S000378
GTGANTG10	212 (-) GTGA	S000378
GTGANTG10	483 (-) GTGA	S000378
GTGANTG10	626 (-) GTGA	S000378
GTGANTG10	639 (-) GTGA	S000378
GTGANTG10	1013 (-) GTGA	S000378
GTGANTG10	1060 (-) GTGA	S000378
GTGANTG10	1282 (-) GTGA	S000378
GTGANTG10	1382 (-) GTGA	S000378
GTGANTG10	1410 (-) GTGA	S000378
GTGANTG10	1550 (-) GTGA	S000378
IBOXCORE	308 (+) GATAA	S000199
IBOXCORE	1349 (+) GATAA	S000199
IBOXCORE	1361 (+) GATAA	S000199
INRNTPSADB	298 (+) YTCANTYY	S000395
INRNTPSADB	194 (+) YTCANTYY	S000395
INRNTPSADB	567 (+) YTCANTYY	S000395
INRNTPSADB	470 (-) YTCANTYY	S000395
INRNTPSADB	690 (-) YTCANTYY	S000395
L1BOXATPDF1	179 (-) TAAATGYA	S000386
L1BOXATPDF1	415 (-) TAAATGYA	S000386
LECPLEACS2	873 (+) TAAAATAT	S000465
LECPLEACS2	861 (-) TAAAATAT	S000465
MARARS	554 (+) WTTTATRTTTW	S000064
MARARS	1432 (+) WTTTATRTTTW	S000064
MARARS	339 (-) WTTTATRTTTW	S000064
MARTBOX	144 (+) TTWTWTTWTT	S000067
MARTBOX	405 (+) TTWTWTTWTT	S000067
MARTBOX	864 (+) TTWTWTTWTT	S000067
MARTBOX	915 (+) TTWTWTTWTT	S000067
MARTBOX	1077 (-) TTWTWTTWTT	S000067
MYB1AT	216 (+) WAACCA	S000408
MYB1AT	253 (-) WAACCA	S000408
MYB2CONSENSUSAT	902 (-) YAACKG	S000409
MYBATRD22	215 (+) CTAACCA	S000175
MYBCORE	902 (+) CNGTTR	S000176
MYBCORE	721 (-) CNGTTR	S000176
MYBGAHV	579 (-) TAACAAA	S000181
MYBGAHV	884 (-) TAACAAA	S000181
MYBST1	512 (+) GGATA	S000180
MYBST1	697 (+) GGATA	S000180
MYBST1	633 (-) GGATA	S000180
MYCATERD1	1411 (-) CATGTG	S000413
MYCATRD22	1411 (+) CACATG	S000174
MYCCONSENSUSAT	600 (+) CANNTG	S000407
MYCCONSENSUSAT	902 (+) CANNTG	S000407
MYCCONSENSUSAT	1173 (+) CANNTG	S000407
MYCCONSENSUSAT	1411 (+) CANNTG	S000407
MYCCONSENSUSAT	600 (-) CANNTG	S000407

MYCCONSENSUSAT	902 (-) CANNTG	S000407
MYCCONSENSUSAT	1173 (-) CANNTG	S000407
MYCCONSENSUSAT	1411 (-) CANNTG	S000407
NODCON1GM	205 (+) AAAGAT	S000461
NODCON2GM	190 (+) CTCTT	S000462
NODCON2GM	1085 (-) CTCTT	S000462
NODCON2GM	1471 (-) CTCTT	S000462
NTBBF1ARROLB	1524 (+) ACTTTA	S000273
NTBBF1ARROLB	440 (-) ACTTTA	S000273
NTBBF1ARROLB	621 (-) ACTTTA	S000273
OSE1ROOTNODULE	205 (+) AAAGAT	S000467
OSE2ROOTNODULE	190 (+) CTCTT	S000468
OSE2ROOTNODULE	1085 (-) CTCTT	S000468
OSE2ROOTNODULE	1471 (-) CTCTT	S000468
POLASIG1	260 (+) AATAAA	S000080
POLASIG1	347 (+) AATAAA	S000080
POLASIG1	1080 (+) AATAAA	S000080
POLASIG1	1128 (+) AATAAA	S000080
POLASIG1	551 (-) AATAAA	S000080
POLASIG1	865 (-) AATAAA	S000080
POLASIG1	916 (-) AATAAA	S000080
POLASIG1	948 (-) AATAAA	S000080
POLASIG1	1540 (-) AATAAA	S000080
POLASIG2	542 (+) AATTAAA	S000081
POLASIG2	1024 (+) AATTAAA	S000081
POLASIG2	1031 (+) AATTAAA	S000081
POLASIG2	373 (-) AATTAAA	S000081
POLASIG2	400 (-) AATTAAA	S000081
POLASIG2	842 (-) AATTAAA	S000081
POLASIG2	1112 (-) AATTAAA	S000081
POLASIG2	1396 (-) AATTAAA	S000081
POLASIG2	1575 (-) AATTAAA	S000081
POLASIG3	539 (+) AATAAT	S000088
POLASIG3	1077 (+) AATAAT	S000088
POLASIG3	38 (-) AATAAT	S000088
POLASIG3	236 (-) AATAAT	S000088
POLASIG3	409 (-) AATAAT	S000088
POLASIG3	868 (-) AATAAT	S000088
POLASIG3	879 (-) AATAAT	S000088
POLASIG3	919 (-) AATAAT	S000088
POLLEN1LELAT52	4 (+) AGAAA	S000245
POLLEN1LELAT52	49 (+) AGAAA	S000245
POLLEN1LELAT52	70 (+) AGAAA	S000245
POLLEN1LELAT52	102 (+) AGAAA	S000245
POLLEN1LELAT52	121 (+) AGAAA	S000245
POLLEN1LELAT52	381 (+) AGAAA	S000245
POLLEN1LELAT52	423 (+) AGAAA	S000245
POLLEN1LELAT52	478 (+) AGAAA	S000245
POLLEN1LELAT52	762 (+) AGAAA	S000245
POLLEN1LELAT52	956 (+) AGAAA	S000245
POLLEN1LELAT52	436 (-) AGAAA	S000245
POLLEN1LELAT52	978 (-) AGAAA	S000245
POLLEN1LELAT52	994 (-) AGAAA	S000245
POLLEN1LELAT52	1066 (-) AGAAA	S000245

POLLEN1LELAT52	1311 (-) AGAAA	S000245
PREATPRODH	1015 (+) ACTCAT	S000450
PYRIMIDINEBOXOSRAMY1A	1308 (+) CCTTTT	S000259
QARBNEXTA	1139 (+) AACGTGT	S000244
QELEMENTZMZM13	1447 (-) AGGTCA	S000254
RAV1AAT	287 (+) CAACA	S000314
RAV1AAT	616 (+) CAACA	S000314
RAV1AAT	646 (+) CAACA	S000314
RAV1AAT	653 (+) CAACA	S000314
RAV1AAT	721 (+) CAACA	S000314
RAV1AAT	1276 (+) CAACA	S000314
RAV1AAT	1297 (+) CAACA	S000314
REALPHALGLHCB21	217 (+) AACCAA	S000362
REALPHALGLHCB21	1205 (-) AACCAA	S000362
RHERPATEXPA7	1260 (+) KCACGW	S000512
RHERPATEXPA7	1060 (+) KCACGW	S000512
ROOTMOTIFTAPOX1	407 (+) ATATT	S000098
ROOTMOTIFTAPOX1	427 (+) ATATT	S000098
ROOTMOTIFTAPOX1	446 (+) ATATT	S000098
ROOTMOTIFTAPOX1	516 (+) ATATT	S000098
ROOTMOTIFTAPOX1	558 (+) ATATT	S000098
ROOTMOTIFTAPOX1	861 (+) ATATT	S000098
ROOTMOTIFTAPOX1	877 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1091 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1286 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1430 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1436 (+) ATATT	S000098
ROOTMOTIFTAPOX1	1592 (+) ATATT	S000098
ROOTMOTIFTAPOX1	341 (-) ATATT	S000098
ROOTMOTIFTAPOX1	393 (-) ATATT	S000098
ROOTMOTIFTAPOX1	426 (-) ATATT	S000098
ROOTMOTIFTAPOX1	496 (-) ATATT	S000098
ROOTMOTIFTAPOX1	876 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1090 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1425 (-) ATATT	S000098
ROOTMOTIFTAPOX1	1591 (-) ATATT	S000098
SEF1MOTIF	1430 (+) ATATTTAWW	S000006
SEF4MOTIFGM7S	518 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	846 (+) RTTTTTR	S000103
SEF4MOTIFGM7S	468 (-) RTTTTTR	S000103
SEF4MOTIFGM7S	1027 (-) RTTTTTR	S000103
SORLIP1AT	1148 (+) GCCAC	S000482
SORLIP1AT	1404 (+) GCCAC	S000482
SP8BFIBSP8BIB	14 (-) TACTATT	S000184
SURECOREATSULT11	111 (+) GAGAC	S000499
SV40COREENHAN	242 (-) GTGGWWHG	S000123
T/GBOXATPIN2	1139 (+) AACGTG	S000458
T/GBOXATPIN2	1061 (-) AACGTG	S000458
T/GBOXATPIN2	1261 (-) AACGTG	S000458
TAAAGSTKST1	440 (+) TAAAG	S000387
TAAAGSTKST1	621 (+) TAAAG	S000387
TAAAGSTKST1	1469 (+) TAAAG	S000387
TAAAGSTKST1	364 (-) TAAAG	S000387

TAAAGSTKST1	1395 (-) TAAAG	S000387
TAAAGSTKST1	1525 (-) TAAAG	S000387
TAAAGSTKST1	1539 (-) TAAAG	S000387
TATABOX2	343 (+) TATAAAT	S000109
TATABOX2	986 (+) TATAAAT	S000109
TATABOX2	554 (-) TATAAAT	S000109
TATABOX2	833 (-) TATAAAT	S000109
TATABOX2	1432 (-) TATAAAT	S000109
TATABOX4	984 (+) TATATAA	S000111
TATABOX5	39 (+) TTATTT	S000203
TATABOX5	410 (+) TTATTT	S000203
TATABOX5	552 (+) TTATTT	S000203
TATABOX5	880 (+) TTATTT	S000203
TATABOX5	949 (+) TTATTT	S000203
TATABOX5	73 (-) TTATTT	S000203
TATABOX5	346 (-) TTATTT	S000203
TATABOX5	538 (-) TTATTT	S000203
TATABOX5	1076 (-) TTATTT	S000203
TATABOXOSPAL	559 (+) TATTTAA	S000400
TATABOXOSPAL	950 (+) TATTTAA	S000400
TATABOXOSPAL	536 (-) TATTTAA	S000400
TATAPVTRNALEU	984 (-) TTTATATA	S000340
TBOXATGAPB	1252 (+) ACTTTG	S000383
TBOXATGAPB	656 (-) ACTTTG	S000383
TRANSINITDICOTS	1514 (-) AMNAUGGC	S000201
TRANSINITMONOCOS	1514 (-) RMNAUGGC	S000202
WBOXATNPR1	1446 (+) TTGAC	S000390
WBOXATNPR1	466 (-) TTGAC	S000390
WBOXHVISO1	624 (-) TGACT	S000442
WBOXNTERF3	1447 (+) TGACY	S000457
WBOXNTERF3	1566 (+) TGACY	S000457
WBOXNTERF3	624 (-) TGACY	S000457
WRKY71OS	1248 (+) TGAC	S000447
WRKY71OS	1345 (+) TGAC	S000447
WRKY71OS	1447 (+) TGAC	S000447
WRKY71OS	1566 (+) TGAC	S000447
WRKY71OS	466 (-) TGAC	S000447
WRKY71OS	625 (-) TGAC	S000447

APPENDIX H.

ALL POSSIBLE PAIRWISE ALIGNMENTS OF FAMILY MEMBER NUCLEOTIDE SEQUENCES

LJFgene3 vs. LJFgene14

```
LJFgene3        ------------------------------------------------------------
LJFgene14       AAGTCCCACCTTCTTTTATTCATCACATGATTCACATCTCATTTCTTATTTTCGGGTCAC


LJFgene3        -------------------TTTCGGTCTGTGAAGATATAT--GTCCATAAGTTCCTTAAT
LJFgene14       TTTGTTAAATTATAAATAATTTCGTTCTGTGAAGGTACACACGTTCATAAGTTCCTTAAT
                           ***** ********* ** *   ** ***************

LJFgene3        TTTCTCGAACCTTCATTTTCAGCTCCCAACAACAATGGCTTCAATGGCATCTTCAAGCTC
LJFgene14       TTTCTCGAACCTTCATTTTCAGCTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTC
                ******************************** ***************************

LJFgene3        CTTCTGCAACCTCAAGTTCATCACCAAACCCAACAATGGTAGAAGAAGCTCTCTTCCCCG
LJFgene14       CTTCTGCAACCTCAAGTTTATCACCAAACCCAACAATGGTAGAAGAAGCTCTCTTCGCCG
                ****************** ************************************ ***

LJFgene3        TATTGTATTCTGTCAGAAGCACCACGATAGCACACCCACCGACCAAATCAACCGAAGGTT
LJFgene14       TATTGTATTTTGTCAGAAGCATCACGATGACACACCCACCGACCAAATCAACCGAAGGTT
                ********* *********** ******  ******************************

LJFgene3        CTTATTTCTTCACACTCGCACTTTCTAATTCCTTTCTATGGATTATTCATATCTATTCAT
LJFgene14       CTTACTTCTTCACACTCACACTTTCTATTTCCTTTCTATTGATTATTCG----------T
                **** ************ ********* *********** ********           *

LJFgene3        ACCCATCTTCTGAAATCTCTTTATATTTCAATTATTTTGTCTATTGAAGAGAACTCATAT
LJFgene14       AACCATCTTCTGAAATCTCGTTACATTTCAATTCTTTTGTGTATTGAAGAGAACTCATAT
                * ***************** *** ********* ****** ********************

LJFgene3        TGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTGAACTTCGGGTACCCCTCCTCTG
LJFgene14       TGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTCAACTTCGGGTACCCCTCCTCTG
                ************************************ ***********************

LJFgene3        CTTGT------------TTTTGGAAAATTTTTGTTTTTCATTTTATTTTGAATGTAAATT
LJFgene14       TTTTTGCTCTGTTTTTTTTTCTGGAAATTTTAGTTTTTCATTTTATTTTGAATGTAAATT
                 ** *            ***  * ******* ****************************

LJFgene3        GAATTCAAGATTTGATTTTGTTGGTGGGTTTGAAGACCCTTTTGGTTTTTAATTTCGGTT
LJFgene14       AAATTCGAGATTTGATTTTGTTAGTGGGTGTTGAGACCCTTTTGGATTTTAGTTTGGGTT
                 ***** *************** ****** *  ************ ***** *** ****

LJFgene3        TTGTTTTGTATTGGACATGGGTGGTGGTTAAAAAAGAGAAAATTGAGTTTGTGTCTTGTG
LJFgene14       GTGTTTTGTATTGGAAATGGGTGGT-----------------------TTGGGTTTTGTG
                 ************** *********                       *** ** *****

LJFgene3        TTTTGATGGTGCAGTGGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCA
LJFgene14       TTTTGGTGGTGCAGTGGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCA
                ***** ******************************************************

LJFgene3        TTAGCTCTGTGCCCGGCAACGAAGAATTGCGTGTCAACCTCTGAGAATATCAGTGATCGC
LJFgene14       TTAGCTCTGTGTCCGCCAACTAAGAACTGCGTGTCAACCTCTGAGAATATCAGCGATCGC
                *********** *** **** ***** ************************ ******

LJFgene3        ACACATTATGCTCCTCCATGGTAAAAGTTTCCTTCTTTTTCTTATTTTAATTTTCACCTT
LJFgene14       ACACATTATGCTCCTCCATGGTAAAAGTTCCCTTCTTTTTCTTATTTTAATTTTCACCTT
                ***************************** ******************************

LJFgene3        GGATTTATGGGATTATATGTATTAAATGCATTTTTTTAATTGTGTGTTTGGACAAACTAC
LJFgene14       GGATTTATGGGATTATATGAATTGAATGCAATTTTTTAATTGTG--TTTGGATAAACAAC
                ****************** *** ****** *************  ****** **** **

LJFgene3        TTAGTTAAGTGCTCATCTCATCATGTAAGTGCTTATGCATAAGTTGTTTCTATAATAAAA
LJFgene14       TTAGTTAACTGCCCAT-----CATGTAAGTGCTTATGTATAAGGTTTTTCTATAGTAAAA
                ******** *** ***     ***************** ***** * ******** *****

LJFgene3        AAATAAAAATACATACGTATGAGTTGTTGTTGTAAGCTTTTTTCTTAAGTTATTCTGGAA
LJFgene14       AAAT------------GTACAA------GTTGCAAGCTTTTTTCTTAATTTATTCTTGAA
                ****            ***  *      **** *************** ******* ***

LJFgene3        ATCTTATTGAAATAATCTGAAAACAACTTTTTTTTTTACATG-ATCTGCTGAAAACAACT
```

```
LJFgene14       ATCTTATTGAAATAATCTGAGAACAACTTTTTCTTTTTTACCTGTGATCTGAAAACAACT
                ******************** *********** ****  *    *   ************

LJFgene3        TATAGACATATGGTAATCACATATCATTAAGTTAAGTTATTCCAAACACTTACATAAACA
LJFgene14       CATAGACATATCATAA--ACTTGGGATATAGATAAATT----------TTTTCATAAACA
                 **********  ***  ** *   **  ** *** **           ** ********

LJFgene3        CTTATAAGAGAAAAACAA-TAAGAAATAAAAACAAAATAAAATTTTCAATAAGTTAAAAT
LJFgene14       CTTACAAGAGAAAAAAAAATACAAAAGAAAAATAAAATAAATTTTTCTATAAGCTAAAAT
                **** ********** ** **  *** ***** ******** ***** ***** ******

LJFgene3        TAGTTTATAAAAGTTTTTTTTTATTATAGAAGCTCTCTTGTATTAGCCTCTCCTAAAGTT
LJFgene14       TAGTGTATGTGTAAGCTAATTTGT------AGTTCTTTCATATTAGCTTCTCCAAAAGTT
                **** ***        *  *** *      ** *** *  ******* ***** ******

LJFgene3        TTTTTTTCTTCATAATTTAGCTTAAAAAGAAACCTATTTCA----TTTTTTCATTTTATT
LJFgene14       TTTTTTTTTTT-TAACTTATGCATAAGTTAAATTTAGCTTAAAGATAAATTTATTTCATT
                ******* **  *** ***     **   ***  **  * *    *   ** **** ***

LJFgene3        TTCTTCTCTTGTAAATGCTTTTGGAGAAGTTTGTCCAAACATACCTTTACACAAATACTT
LJFgene14       TTCTTCTCTTGTAAATGCTTTTGGAGAAATGTATCCAAACAGACCTTTACACAAGTACTA
                ***************************** * * ******** ************ ****

LJFgene3        ATAAGATAAGTCTAATTAAGCTTTTTCAAACACACTCAAAGTTAAAGTATTTCCATTTTG
LJFgene14       ATAAGATAAGTCTAATTAAGCCTTTTCAAA-ACGCTCAAAGTTCAAGTATTTCCATTATG
                ********************** ******** ** ********* ************* **

LJFgene3        TTGTTTTTTTGGGCCTTAATTTTCGGTTAACTAACTTGTGGTGTCTAAATTTCGTCTTGA
LJFgene14       TTGATTCTTCGGGCCTTAATTTTCGGTTAAGTAACTTGTGGTGTCAAAATTGCGTGTTGA
                *** ** ** ******************** ************** ***** *** ****

LJFgene3        GGAGGGTCTGATAAACCAGATTTAGGAGATAAGTAAGCAGTCAGTGTTTATTTATTTATT
LJFgene14       GGAGGGTCTGGTAAACCAGATTTAGGAGATAAGTAA----TCAGGGTTTATTTATTTAT-
                ********** ************************    **** **************

LJFgene3        TTTTGGATTTAGTCCAAAAGGAACCTATGGC-ATATTTGTGAGAATCACGTTGCAATA--
LJFgene14       ---TGGATTTAGTCCAAACGGAACCTATGGCCATATTTGTGAGAATCACGTTGCAATAAT
                   *************** ************ *************************

LJFgene3        -GACAATAGTGCACCTGGAACGATTTATCACGTTTTAAAC-------AAGTAGTGCAACC
LJFgene14       AGACAATGGTGCACTTGGAACAATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAGCC
                 ****** ****** ****** ******************       ********** **

LJFgene3        GATATTGGACTATTGACTGTTGAACATTGTTGACTTGACATAAGAATTGGACAATTGGTC
LJFgene14       GAAATTGGACTATTGACTGTTGAACAATGTTGACTTGACAT--GAATTGGACTATTGGTC
                ** ********************** **************  ********* *******

LJFgene3        ACACACACATTGGCCGGTGAAAGTAGTGCAATCTTTACTTTTTATTTCTTTTT-GACAAA
LJFgene14       ACACACA--TTGGCCGGTGAAAGCAGTGCAATCTTAACTTTTTCTTTTTTTTTTTGACAAC
                *******  ************** ************ ******* *** ***** *****

LJFgene3        AAAATTTT---CTCTACCTTTGGCCCTTGT-TTGAGCATGGTCGCCACAAAATCCAAACT
LJFgene14       TTTTTTTTTTTCTCTATCTTTGGTCCTTGTATTGAGCATGGTCCCACCAAAATCCAAACT
                    ****   ***** ****** ****** ************ *  **************

LJFgene3        TTATTATTG--TATAGATGAATCATGATCTCACTTTGTTTTAGTATTTTCATTCTTTTTC
LJFgene14       TTATTATTGGACATAGATGAATCATGATGTCACTTTGTTTTAATTTTTAGTTTCTTTTTC
                *********   ***************** ************* * ***   *********

LJFgene3        AGGAACTATAATCCTGAAGGTAGGAAAAAACCTGTGAACAGAGAGGAAGCAATGGAGGAA
LJFgene14       AGGAACTATAATCCTGAAGGAAGGAAAAAACCTGTGAGCAGAGAGGAAGCAATGGAGGAA
                ******************** ***************** **********************

LJFgene3        CTGATAGACGTGGTAATAAATCTAACTGAACTGAAATTTTGAATTATATCACTGGATTGC
LJFgene14       CTGATAGACGTGGTAATAAATCTAGCTGAACTGAAATCTTGAGTTATAACACTAGATTGC
                ************************ ************* **** ***** **** ******

LJFgene3        AATTTTCTTTTCCTTCCCTCTTTTAATCAACATCGATTATAATTTATAAAATTTATAAAA
LJFgene14       AATTTTCTTTTCCTTCCCTCATTTTATCAACATCGATTATAATTTATAAAATTTATAAAA
                ******************** *** ************************************
```

```
LJFgene3        GGGGTGATAAACTGAAAGTAAATATACACTTTGTAATGCACTATTCCTAAC---ACACTT
LJFgene14       GGAGTGATA--------GTAAATATACACTTTGTAACACACTATTCTTAACCCCACCCTT
                ** ******        ******************  ******** ****   ** ***

LJFgene3        TCTATTATTCATTAAAATTTATTGAAAATTACGAAGTCATGGGTGGAATTCATTGAATAA
LJFgene14       TTTATTATTGGTTAAAATTTATCAGAAATTAGAAAGTCATGAGTGGAACTCATTGAATAA
                * *******  ***********   ******  ******** ****** ***********

LJFgene3        AGAGTAAGACCTACATGATTTTGTAATTTCTAATAAACTCTTACTATTAATAAAGAATGT
LJFgene14       AGAGTGAGACCTATGTGATTTTATAATTTCTAATAAACT--TATTATTAATAG----TGT
                ***** *******  ******* ****************  ** ********     ***

LJFgene3        ATTTAAAAGGGTATGTTGTCAACATTTCTCTAATTTATAGTTGATTTGTGATAATAGCTG
LJFgene14       ATTTAAAAGGATACATTGTGAACATTTCTCTAATTTATAATTGATTTGTAATAATAGCTG
                ********** **  **** ****************** ********* **********

LJFgene3        GTTATTTGCACTTTTCCTTCCAATCATTGAAGTAAAGTTAAA-CCAGTTCCGGACAATTT
LJFgene14       GTTATTTGCACTTTTCC----AATCATTGAAGTAAAGTTAAAACTAGTTCCGGATAATTT
                *****************    ********************* * ********* *****

LJFgene3        TGCAGATAGAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAG
LJFgene14       TACAGATCGAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAG
                * ***** ****************************************************

LJFgene3        ACTATATTCGTGTGGAGTACCAAAGCTCAATTTTGGGGGTAAGTGTAACTTACATCTAAG
LJFgene14       ACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGGTAAGTGTAACTTACATCTAAG
                ******************************* ****************************

LJFgene3        GAAACTCATCACGAAGAAAAATGATAATTTTATACATTTGAGATGATATCAAGATTCAAG
LJFgene14       GAAACTCATCTTGAAGAAAAATGATCATTTTATACATTTGAGATGATATCAAGA------
                **********  ************* ****************************

LJFgene3        AACCATCTCTAATTTCCTTCCCTCCTTTTTCTGTCATGTGCTAACAGTTTGTAGATGATG
LJFgene14       -ACCATCTCTAATTTCCTTCTCCCTTTTTTCTGTCATGTGCTAACAGTTTGTGGATGATG
                 ******************* * * ***************************** *******

LJFgene3        TTGAGTTCTGGTTCCCACCGGGTAAGGGTTCTACTGTGGAGTACCGATCTGCATCTCGGT
LJFgene14       TTGAGTTCTGGTTTCCACCCGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTCGGT
                ************* ***** ********************** ****************

LJFgene3        TAGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTGTGATTTCATAATT--C
LJFgene14       TGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATGATTTCATAATTAAT
                * ************************************** *************

LJFgene3        ATGTGTTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCATGGTAAAACTCCTCTTTC
LJFgene14       ATGTGCTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCAGGGTAAAACTCC-CCTTC
                ***** ********************************** *********** * ***

LJFgene3        CTTCATGTCATGTCAAACATTTTATACTCAAGTAGATGATTCACTAAATTTGAGTCTCAA
LJFgene14       CTTCATGTCATGTCAAACATTTTATACTTAAGTAG---ATTCACTAAATTTGAGTCTCAA
                **************************** ******   **********************

LJFgene3        ATGTTTTAACTTTATTCTAAATTAGTCACTTATTTTAACTGAAGGTAAATTTGGTTAACT
LJFgene14       ATGTTTTAACTTTATTCTAAATTAGTCACTTATTTTAACGGAAGGTAAATTTGGTTGACT
                ************************************ **************** ***

LJFgene3        ATGATCAGAAATACATTGACATTTTTTAATTGGTAGAGATAAAGAATATTTTTTATGTAC
LJFgene14       ATGATGAGAAATACGTTGATATTTTTTAATTGGTAGAGATAAAGAATATTTTTTATGTAC
                ***** ******** **** ****************************************

LJFgene3        AATAAAGAGAGTATTTACTCCAGAGGATGTAAATCCCTTGCTAAATATTTTTGTGATGAA
LJFgene14       AATAAAGAGAGTATTTACTCCAGAGGATGCAAATCCCTTACTAAATATTTTTGTGATGAA
                **************************** ********* ********************

LJFgene3        AAATCTTGGTTGTTGACAGGCACTGCGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCA
LJFgene14       AAATCTTGGTTGCTGACAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTCA
                ************ ************ ************************** *******

LJFgene3        AGACACCATATGATGAATAAACTCAGGCAGAATTAACATCAGCATCTAAGCAAATATTAT
LJFgene14       AGATACCATATGATTAATAAACTCAGGCTGAATTAGCATCAGCATCTAAGCAAATATTAT
                *** ********** ************* ****** ************************
```

```
LJFgene3        TTCATATACTTTGTGACCTTGTATACATTTGTATTAGATACAAATCTCACAGGATCATTG
LJFgene14       TTCATATACTTTG-GACCTTGTATACTTTTGTATTAGATACAAATCGCACAGGATCATTG
                ************* ************ ****************** *************

LJFgene3        AAAGCAAACTTTTCTTTGATTATTGGAATTGTAGAGAAATCATTGAGAACAGTACTTCAA
LJFgene14       CAAGCAAACTTTTCTTAGATTTTTGGAATTGTAGAGAAATCATTGAGAACAGTACTTCAA
                 *************** **** ************************************

LJFgene3        ACTCTC--GGGGAAGGAATGAAATGAAGACCTTGCCCCATATCCTTCTTCAAGTTCATTA
LJFgene14       ACTCTCTCGGGGAAGGAATGAAATGAAGACCTTGCGCCATATC-TTCTTCAAGTTCATTA
                ******  *************************** ******* ****************

LJFgene3        ATTGGTCCGCTTATTTACTCTTTCACCAAGTTCAATCTAACAATGTATCGTTCTTGTTTC
LJFgene14       ATTGGTCCACTTATTTACACCTTCACCGAGTTCAATCTAATAATGTATGAATCTTGTTTC
                ******** ********* * ****** ************ *******   *********

LJFgene3        AAGAATATTAATTTGTGTTTTGTTTCTAATGTGTTCTTCACGATTCACTTTTGTTGTAAG
LJFgene14       AAGAAATTTAATTTGTGTTTTGTTTCAAACGTGTTCTTCATGATTCACTTTTGTTGTAAG
                *****  ******************* ** ********** *******************

LJFgene3        TTTGACTTAGCTTCTTCATGTATTAAAACTTAACCTTGTGA--------------------
LJFgene14       TTTGACTTTGCTTCTTCATGTATTAAAGCTTATCCTTGCGATCAATTTGGATGATGCTTT
                ******** ****************** **** ***** **

LJFgene3        ------------------------------------------------------------
LJFgene14       AGGCCTTTCTGTCATCAAAATGTACGTACCTATGTCATGTTTTCAGAGGCTTTTTGATAT


LJFgene3        ----------------------
LJFgene14       CCTTGATGTCTTTACAAGAAA
```

LJFgene3 vs. LJFgene1

```
LJFgene3        --------------------CCATAAAAAAGA-------AAAAAAAAAAGTCCCACCGCCC
LJFgene1        TAAAAAGGAGAAAGAAAAAACAAAAAAAAAAATGTTGCGGAAACGAAGCGTCCCACCACCC
                                   ** ****** *        ***  **  ******** ***

LJFgene3        ACCTTCTTT-----------ATCACATGATTCACATCTCATTCCTTATATTTGGTTCACA
LJFgene1        ACCTTCTTTTTATATTCATCATTACATGATTCACATCTCATTCCATATTTTCGGGTCACA
                *********           ** ********************* *** ** ** *****

LJFgene3        TTCTTAAATTAT---AAATATTTCGGTC-TGTGAAGATATATGTCCATAAGTTCCTTAAT
LJFgene1        TTCTCAAATTATTATAACTAATTTCGTCATGTGAAGATAC--GTTCATAAGTTCCTTAAT
                **** *******   ** ** **  *** **********   ** ***************

LJFgene3        TTTCTCGAACCTTCATTTTCAGCTCCCAACAACAATGGCTTCAATGGCATCTTCAAGCTC
LJFgene1        TTTCTTGAACCTTTATTTTCAGCTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTC
                ***** ******* ****************** ***************************

LJFgene3        CTTCTGCAACCTCAAGTTCATCACCAAACCCAACAATGGTAGAAGAAGC---------TC
LJFgene1        CTTCTGCAACCTCAAGTTTATCACCAAACCCAACAACAATGGTAGAACCAATGCTTCTTC
                ****************** ****************   * * **** *         **

LJFgene3        TCTTCCCCGTATTGTATTCTGTCAGAAGCACCACGATAGCACACCCACCGACCAAATCAA
LJFgene1        TCTTCCCCGTATTGTATTCTGTCAGAAGCACAACGATGACACCCCCACCGACCAAATCAA
                ******************************** *****  *** *****************

LJFgene3        CCGAAGGTTCTTATTTCTTCACACTCGCACTTTCTAATTCCTT---TCTATGGATTATTC
LJFgene1        CCGAAGGTTCTTACTTCTTCACACTCACACTCTCTCACTCCTTCTTTCTATTGATTATTT
                ************* ************ **** *** * *****   ***** *******

LJFgene3        ATATCTATTCATACCCATCTTCTGAAATCTCTTTATATTTCAATTATTTT-GTCTATTGA
LJFgene1        ATAACTATTCATACCCATCTTCTGAAATCTCTTTACATTTCAATTCTTTTTGTGTATTGA
                *** ******************************** ********* **** ** ******

LJFgene3        AGAGAACTCATATTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTGAACTTCGGG
LJFgene1        AGAGAACTCATATTGAGAAGCAGTGAAATAGCGACCATTGGTGCCATCTTCAACTTCGGG
                ********************** ************************* *********

LJFgene3        TACCCCTCCTCTGCTTGT----------------TTTTGGAAAATTTTTGTTTTTCATTT
```

```
LJFgene1        TACCCCTCCTCTGTTTTTCCTCGGTTTTTTTCTTTTTTGGAAAATTTAGTTTTTTCATTT
                ************* ** *                *************   **********

LJFgene3        TATTTTGAATGTAAATTGAATTCAAGATTTGATTTTGTTGGTGGGTTTGAAGACCCTTTT
LJFgene1        TATTTTGAATGTAAATTGTATTCAAGATTTGATTTTGTTGGTGGGTTTGGAGACCCTTTT
                ****************** ****************************** **********

LJFgene3        GGTTTTTAATTTCGGTTTTGTTTTGTATTGGACATGGGTGGTGGTTAAAAAAGAGAAAAT
LJFgene1        GGATTTTAGTTTCAGTTTTGTATTGTATTGGAAATGGGTGGTGGTTAAAAAAGAGAAAAT
                ** ***** **** ******* ********** ****************************

LJFgene3        TGAGTTTGTGTCTTGTGTTTTGATGGTGCAGTGGGAAAAAACCTGATTATCTTGGAGTGC
LJFgene1        TGAGTTTGGGTTTTGTGTTTTGGTGGTGCAGTGGGAAAAAACCTGATTATCTTGGAGTGC
                ******** ** ********** ************************************

LJFgene3        AGAAAAACCCACCAGCATTAGCTCTGTGCCCGGCAACGAAGAATTGCGTGTCAACCTCTG
LJFgene1        AGAAAAACCCACCAGCATTAGCTCTGTGTCCGGCAACTAAGAACTGCGTGTCAACCTCTG
                **************************** ******** ***** ****************

LJFgene3        AGAATATCAGTGATCGCACACATTATGCTCCTCCATGGTAAAAGTTTCCTTCTTTTTCTT
LJFgene1        AGAATATCAGTGATCGCACACATTATGCTCCTCCATGGTTAAAGTTCCCCTCTTTTTCTT
                ************************************ ****** ** **********

LJFgene3        ATTTTAATTTTCACCTTGGATTTATGGGATTATATGTATTAAATGCA--TTTTTTTAATT
LJFgene1        ATTTTAATTTTCACCTTCGATTTATGGGATTATATGAATTAAATGCAATTTTTTTTACTT
                ***************** ****************** **********  ******** **

LJFgene3        GTGTGTTTGGACAAACTACTTAGTTAAGTGCTCATCTCATCATGTAAGTGCTTATGCATA
LJFgene1        GTCTGTTTGGATAAACAACTTAGTTAAGTATTCA-----TCATGTAAGTGCTTATGTATA
                ** ******** **** ************  ***     ***************** ***

LJFgene3        AGTTGTTTCTATAATAAA--AAAAT-----------------------------------
LJFgene1        AGTTGTTTCTATAATAAATAAAAATGAGGAGGGAAAGTTATTCCAAAAATCACTTTAAAA
                ******************  *****

LJFgene3        ------------------------------------------------------------
LJFgene1        GAGGTACTCACTTTATTTTAATTATTGATTTTTTTAAAATTTAATGATTAAGATTAATTA


LJFgene3        -------------------------AAAAATACATA--------------------------
LJFgene1        TTTATTATAATTATTAGATTCTAAAAAAATAAATAATAAGGATAATAAATAACTCTAAAA
                                        ******* ***

LJFgene3        ----------------------------------------CGTATGAGTTGTTGTTGTAA
LJFgene1        AAATTATTCTTAGAGCAAGTTGATGTTTTCCTAAAAAAAATATATAAATTGTTTATATAA
                                                    *** * *****  * ***

LJFgene3        GCT-----TTTTTCTTAAGTTATTCTGGAAATCTTATTGAAATAATCTGAAAACAACTTT
LJFgene1        GTCGTAAGCTTTTCGGAAATTATTCTTGAAATCTTATTGAAATAATCTGAAAACAACTTT
                *        *****  ** ******* **********************************

LJFgene3        TTTTTTTACATGATCTGCTGAAAACAACTTATAGACATATGGTAATCACATATCATTAAG
LJFgene1        TTTT--TACATGATTTG---AAAACAACTTATAGACATATCATAATCACATATCATTAAG
                ****  ******** **   ********************  ******************

LJFgene3        TTA--------AGTTATTCCAAACACTTACAT-AAACACTTATAAGAGAAAAACAATAAG
LJFgene1        TTATTTTATTAAGTCATTTTCATAATTTATTTCAAACACTTACATAAATACTTATAAGAG
                ***        *** ***   *  * ***  * ********* *  *  *     *  **

LJFgene3        AAATAAAAACAAAATAAAATTTTCAATAAGTTAAAATTAGTTTATAAAAGTTTTTTTTTA
LJFgene1        AAAATAAAATAAAATTAATTTCTTTATAAGCTATAA--AATTAGTTAATGTATAAGTCAA
                ***  **** ***** ** ** *  ***** ** **  * **  * ** ** *   *  *

LJFgene3        TTA-TAGAAGCTCTCTTGTATTAGCCTCTCCTAAAGTTTTTTTTTC--TTCATA------
LJFgene1        TAGGTAGAAGCTCTCTCGTATTAAC-TCTTCAAAAGTTTATTTTTTAACTTATATATAAG
                *   ************ ****** * *** * ******* *****    * ***

LJFgene3        ----ATTTAGCTTAAAA-AGAAACCTATTTCATTTTTTCATTTTATTTTCTTCTCTTGTA
LJFgene1        CTAAATTTAACTTAAAAGAGAAACTTATTTCATTTTTTC-TCTTCCTTTCTTTTCTAGTA
                    ***** ******* ****** ************** * **  ****** *** ***
```

```
LJFgene3        AATGCTTTTGGAGAAGTTTGTCCAAACATACCTTTACACAAATACTTATAAGATAAGTCT
LJFgene1        AATGTTTTTATAAAAGTTTACCCAAACAGAACTTTACACAAGTACTTATAAGATAAGTCT
                **** ****  * ******  ******* * ********** ******************

LJFgene3        AATTAAGCTTTTTCAAACACACTCAAAGTTAAAGTATTTCCATTTTGTTGTTTTTTTGGG
LJFgene1        AATTAAGCTTTTTCAAACATGCTCAAAGTTAAAGTGTTTCCA---TAATGTTTTTTTGGG
                ******************  ************** ******   *  ************

LJFgene3        CCTTAATTTTCGGTTAACTAACTTGTGGTGTCTAAATTTCGTCTTGAGGAGGGTCTGATA
LJFgene1        CCTTAATTTTCG-TCAAGTAACTTGTGATGTCAAAATTGCGTGTTGAGGAGGGTCTGGTC
                ************ * ** ********* **** ***** *** ************** *

LJFgene3        AACCAGATTTAGGAGATAAGTAAGCAGTCAGTGTTTATTTATTTATTTTTTGGATTTAGT
LJFgene1        AACTAGATTTAGGAGATAATTAAGCAGTCAGGGTTTATTTATTTATT----GGATTTAAT
                *** *************** *********** ***************    ******* *

LJFgene3        CCAAAAGGAACCTATGGCATATTTGTGAGAATCACGTTGCAATAGACAATAGTGCACCTG
LJFgene1        CCAAAAGGAACCTATGACATATTTGTGAGAATCACGTTGCAATAGATAATGGTGCACTTG
                **************** ***************************** *** ****** **

LJFgene3        GAACGATTTATCACGTTTTAAAC-------AAGTAGTGCAACCGATATTGGACTATTGAC
LJFgene1        GAACAATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAACCGATATTGGACTATTGAC
                **** ******************       ******************************

LJFgene3        TGTTGAACATTGTTGACTTGACATAAGAATTGGACAATTGGTCACACACACATTGGCCGG
LJFgene1        TGTTGAACATTGTTGACTTGACATAAGAAATGAACTATTGGTCACACACGCATTGGCCGG
                **************************** ** ** ************* **********

LJFgene3        TGAAAGTAGTGCAATCTTTACTTTTTATTTCTTTTTGACAAAAAAATTTTCTCTACCTTT
LJFgene1        TGAAAGTAGTGCAATCTTTACTTTTTCTTT--TTTTGAAATTTTTTTTTTCACTACCTTT
                ************************** ***  ****** *      ***** ********

LJFgene3        GGCCCTTGT-TTGAGCATGGTCGCCACAAAATCCAAACTTTATTATTG--TATAGATGAA
LJFgene1        GGTCCTTGTATTGAGCATGGTCCCACCAAAATCCAAACTTTATTATTGGACATAGATGAA
                ** ****** ************ *  ***********************   *********

LJFgene3        TCATGATCTCACTTTGTTTTAGTATTTTCATTCTTTTTCAGGAACTATAATCCTGAAGGT
LJFgene1        TCATGATGTCACTTTGTTTTAATATTTTCATTCTTTTTCAGGAACTATAATCCTGAAGGT
                ******* ************* ***************************************

LJFgene3        AGGAAAAAACCTGTGAACAGAGAGGAAGCAATGGAGGAACTGATAGACGTGGTAATAAAT
LJFgene1        AGGAAAAAACCTGTGAGCAGGGAAGAAGCAATGGAGGAACTTATAGACGTGGTAATAAAT
                **************** *** ** ***************** ******************

LJFgene3        CTAACTGAACTGAAATTTTGAATTATATCACTGGATTGCAATTTTCTTTTCCTTCCCTCT
LJFgene1        GCAACTGAACTGAAATCTTGAGTTA--TCACTGGATTGAAATTTTCTTTTCCTTCCCTCA
                  ************** **** ***  *********** ********************

LJFgene3        TTTAATCAACATCGATTATAATTTATAAAATTTATAAAAGGGGTGATAA-----------
LJFgene1        TTTTATCAACATTGATTATAATTTATAAAATTTATGAAAGGAGTGATAGTGAATATACAC
                *** ******** ********************** ***** ******

LJFgene3        ----------------------------------------------ACTGAAAG------
LJFgene1        TTTGTAACACTATTTCTAATACACTTTCTATTATCGGTTAAAATTTATTGAAAACAGTGT
                                                             * *****

LJFgene3        ---TAAATA-----------------------------------------------------
LJFgene1        TGTTAAATGGCGGCCATGGCGGCGCCATGGCGGAGTTGCGTAACGGTTTTCTGAAAAAAC
                   *****

LJFgene3        ------------------------------------------------------------
LJFgene1        GCCACCGAATAACGGTGGCGTGGCGGATTAAAGATGGCGGCGCCATGGCGGCCGCCATAG


LJFgene3        ------------------------------------------------------------
LJFgene1        CCATGGCGGCCATGGCGGATGTGGCGGGGAGGCGGAAAATGGCAGAATTTTTTTTTTTGT


LJFgene3        ------------------------------------------------------------
LJFgene1        CCGCGGTAGGAGTTGGGCTGACCCGATCCAACCCTACCCGAAAACTTAATGAAAACCATG
```

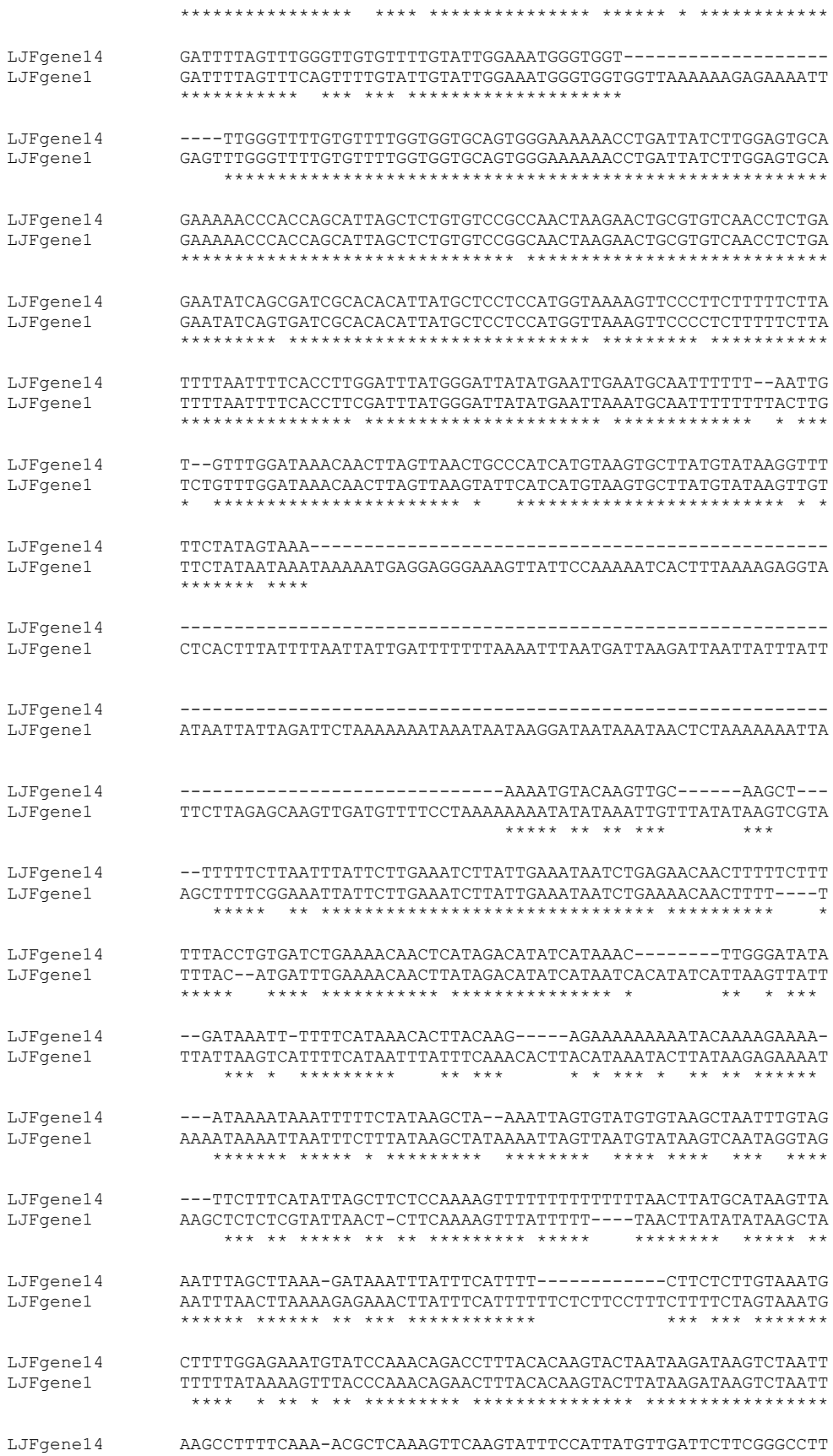
```
LJFgene3        --------TACACTTTGTAATGC---------------------------------------
LJFgene1        ACCCCCCCTACCTTTCAGAACGCTGCTGCAACCCTCGAAGCTTCAACATAGCGCGACCTC
                        ***  **   ** **

LJFgene3        ----------ACTATTCCTA-----------------------------------ACACA
LJFgene1        GGAACCAGCCACGACCCCTGCAACCAGCCTCGGAACCAGCCGCGCTTGACCCGCTGCACA
                         ** *  ***                                      ****

LJFgene3        C-----------------------------------------------------------
LJFgene1        CAGCAGCCAGCAGCGACGACCCATGGCGTGAACGGCGGCGACGAGCACGGCGTGAACAAC
                *

LJFgene3        -----------------------------------------------------------
LJFgene1        GGCGACGCGAACGGAGAAGACGACCCAGAACCGCGACCTCGACTTATACGGGTGGGTGGG


LJFgene3        -----------------------------------------------------------
LJFgene1        CCAAAGCCTTTTTTTTGTTTTTGCTGCTTGACACCCCTTTTTTATGTCTGCAGCTTTTTT


LJFgene3        ------TTTCTATT-----------------------------------------------
LJFgene1        TCCGTTTTTCTATTTGACACCCCCTTTTACTTTCGACAGTCCCATTTTTAATTTTTTTTT
                      ********

LJFgene3        -----------------------------------------------------------
LJFgene1        CTATTTGACACCACAATTTTTTTTTCTGTTCAGTCCTCTTTTAAATGGCTGAACCATCAT


LJFgene3        ------------------------------------------------ATTCATTAAAATTT
LJFgene1        CCTCTTTAATGAGTTTATTTGGTGTTGGTACTTGATTATTGTATGAACTCATGAAACTTT
                                                               * **** *** ***

LJFgene3        -----------------------------------------------------------
LJFgene1        TTTAGTTTATTTGAATGCAATCCTTTGTTTTTTTCAATTTCAATGAGTTTATATATATGT


LJFgene3        -----------------------------------------------------------
LJFgene1        TTTTTTTTTTTTTTGGTCCGCCATGACTTCCGCCATTTTCCGCTACGCCATCCGCCATAT


LJFgene3        ------------------------------------------------------------AT
LJFgene1        TTTTATGGCGGATTTTTTACTTTCCGCCATGAACCGCCATCCGCCATTAACAACATTGAT
                                                                          **

LJFgene3        TGAAAATTACGAAGTCATGGGTGGAATTCATTGAATAAAGAGTAAGA-CCTACATGATTT
LJFgene1        TGAAAATTATGAAGTCATGAGAGGAGCTCATTGAATAAAGAGTGAGAACTTACATGATTT
                ********* ********* * ***  *************** *** * **********

LJFgene3        TGTAATTTCTAATAAACTCTTACTATTAATAAAGAATGTATTTAAAAGGGTATGTTGTC-
LJFgene1        TGTAATTTCTAATAATTTTTTACTATTAATAAAGAGTGTATTTAAATGGGTATGTTTTTG
                ***************  * ***************** ********** ********* *

LJFgene3        AACATTTCTCTAATTTATAGTTGATTTGTGATAATAGCTGGTTATTTGCACTTTTCCTTC
LJFgene1        AACATTTTTCTAATTTCTAATCGATTTGTAATAATAGCTGGTTATTTGCACTTT----CC
                ******* ******** ** * ******* ***********************     *

LJFgene3        CAATCATTGAAGTAAAGTTAAACCAGTTCCGGACAATTTTGCAGATAGAATCAACAACAC
LJFgene1        CAATCATTGAAGTAAAGTTAAACCAGTTCCGGATAATTTTACAGATAGAATCAACAACAC
                ******************************** ****** ******************

LJFgene3        CAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAGTACC
LJFgene1        CAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAGTACC
                ************************************************************

LJFgene3        AAAGCTCAATTTTGGGGGTAAGTGTAACTTACATCTAAGGAAACTCATCACGAAGAAAAA
LJFgene1        AAAGCTCAATCTTGGGGGTAAGTGTAACTTACATCTAAGGAAACTCATCATGAAGAAAAA
                ********** ************************************** *********

LJFgene3        TGATAATTTTATACATTTGAGATGATATCAAGATTCAAGAACCATCTCTAATTTCCTTCC
LJFgene1        TTATCCTTTTATACATTTTAGATGATATCAAGATTCAAGAACCATCTTTAATTTCCTTCT
```

```
                * **  ************ ***************************** ************

LJFgene3        CTCCTTTTTCTGTCATGTGCTAACAGTTTGTAGATGATGTTGAGTTCTGGTTCCCACCGG
LJFgene1        CCCTTTTTTCTGTCATGTGCTAACAGTTTGTGGATGATGTTGAGTTCTGGTTTCCTCCGG
                * * ************************** ******************** ** ****

LJFgene3        GTAAGGGTTCTACTGTGGAGTACCGATCTGCATCTCGGTTAGGAAACTTTGATTTTGATG
LJFgene1        GTAAGGGTTCTACTGTGGAGTATCGTTCTGCATCTCGGTTGGGAAACTTTGATTTTGATG
                ********************** ** ************** ********************

LJFgene3        TGAACAGAAAAAGAATAAAGGTGTGATTTCATAATTCATGTGT--TTTCTCTATAGTTAG
LJFgene1        TGAACAGAAAAAGAATAAAGGTATGATTCCATAATTCATATGTGCTTTCTCTATAGTTAG
                ********************** ***** ********** ***  ****************

LJFgene3        ATAAAGAAATTCTTGGTTCCATGGTAAAACTCCTCTTTCCTTCATGTCATGTCAAACATT
LJFgene1        ATAAAGAAATTCTTGGTTCCAGGGTAAAACTCCCCTTTCCTTCATGTCATGTGAAGCATT
                ********************* *********** ****************** ** ****

LJFgene3        TTATACTCAAGTAGATGATTCACTAAATTTGAGTCTCAAATGTTTTAACTTTATTCTAAA
LJFgene1        TTTTACTCAAGTAGAT---CCACTAAATTTGAGTCTCAAATGTTTTAACTTTATTCTAAA
                ** *************    ****************************************

LJFgene3        T----TAG-------------TCACTTATTTTAACTGAAGGTAAATTTGGTTAACTATGA
LJFgene1        TGTTTTAACTTTATTTGAGTCTCAAATGTTTTAACTGAAGGTAAATTTGGTTAACTATGA
                *    **              ***  * ********************************

LJFgene3        TCAGAAATACATTGACATTTTTTAATTGGTAGAGATAAAGAATATTTTTTATGTACAATA
LJFgene1        TCAGAAATACATTAACA-------------AGAGTTGAAGAATATTTTTTATGTACAATA
                ************* ***             **** * ************************

LJFgene3        AAGAGAGTATTTACTCCAGAGGATGTAAATCCCTTGCTAAATATTTTTGTGATGAAAAAT
LJFgene1        AAGAGAGTATTTGCTCGAGAGAATGTAAATCCTTTTCTAAATATTTTTGTGATGAAAAAT
                ************ *** **** ********** ** ************************

LJFgene3        CTTGGTTGTTGACAGGCACTGCGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCAAGAC
LJFgene1        AATGGTTGCTGGCAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCAAGAC
                  ****** ** ********* ***************************************

LJFgene3        ACCATATGATGAATAAACTCAGGCAGAATTAACATCAGCATCTAAGCAAATATTATTTCA
LJFgene1        ACCATATGATGAAAAAACTTAGGCAGAATTCACATCAGCATCTAAGAAAATATTGTTTCA
                ************* ***** ********** *************** ******* *****

LJFgene3        TATACTTTGTGACCTTGTATACATTTGTATTAGATACAAA-TCTCACAGGATCATTGAAA
LJFgene1        TATACATTGTAACCTTGTATACTTTTGTATTAGATACAAAATCTCACAAGATCATTGAAA
                ***** **** *********** ***************** ******* ***********

LJFgene3        GCAAACTTTTCTTTGATTATTGGAATTGTAGAGAAATCATTGAGAACAGTACTTCAAACT
LJFgene1        GCAAACTCTTCAT-GATTATTGGAATTGTAGA--AATGATTGAGAACAGTACTTCAAACT
                ******* *** * ******************  *** **********************

LJFgene3        CTCGGGG-AAGGAATGAAATGAAGACCTTGCCCCATATCCTTCTTCAAGTTCATTAATTG
LJFgene1        CTCGGGGGAAGGAATGAAATGAAGATGTTACCC-ATATCTTT-TTGAACTTCATTAATTG
                ******* *****************  ** *** ***** ** ** ** ***********

LJFgene3        GTCCGCTTATTTACTCTTTCACCAAGTTCAATCTAACAATGTATCGTTCTTGTTTCAAGA
LJFgene1        GTCCACTTATTTACTCTTTCGCTGAGTTCAATCTAACAATGTAGCATTCTTGTTTCAAGA
                **** *************** *  ****************** * **************

LJFgene3        ATATTAATTTGTGTTTTGTTTCTAATGTGTTCTTCACGATTCACTTTTGTTGTAAGTTTG
LJFgene1        ATTTTATGTTGTGTTGTTTTTCAAATGCATTTGCCATGATTCACCTTCGGTGCATATAGC
                ** ***  ******* * **** ****  **   ** ******* ** * ** *  *

LJFgene3        ACTTAGCTTCTTCATGTATTAAAACTTAACCTTGTGATCAAT--TTGGATACTTTAGGCC
LJFgene1        ATTTGGAATCTGATTATAATAGGAGC-ATTGCGGAGACCAAAAACAGGGTGCATT-----
                * ** *  ***   * ** **  *   *     * ** ***     ** * * **

LJFgene3        TTTCTCTTATCAAAATGTACCTATGTCATGATCTCGAAATTGTTCAATCTCATTGATCCT
LJFgene1        ------------------------------------------------------------


LJFgene3        AACTAGAAAATTCGTCCAA
```

```
LJFgene1       --------------------
```

LJFgene1 vs. LJFgene14

```
LJFgene14      TATATATACACACACACACGAATAAACAAATTTTTTTAAGAGTAAATTACATAACATCTT
LJFgene1       ------------------------------------------------------------


LJFgene14      GTGAGATTTTAAATTTTTTTATACATATTTAAAAAAAAAAGACTTACACAAATCTATCAA
LJFgene1       ------------------------------------------------------------


LJFgene14      TTAAATTTTAAAAAATTACACACGTCTCATAACTGTTTTTGAATAAATACTAACTAAAAT
LJFgene1       ------------------------------------------------------------


LJFgene14      TAAAAAAAATGTAGAAATGCATTATTATTTTTACCGAGTAAAAACATTCTTGATGCGCGA
LJFgene1       ------------------------------------------------------------


LJFgene14      ATTTGACAAAAACCTTTTCGTACAGATAAGCATTTATGGATATTTTTAGTATCCAAAATT
LJFgene1       ------------------------------------------------------------


LJFgene14      GTCACTTTCTCAAACAATCGAAATATATACTATTTATTTTCTAAATATCTAAATCCATAA
LJFgene1       ------------------------------------------------------------TAA
                                                                           ***

LJFgene14      AAAGGAAAATAAATAAA-AAAATAAAAATGTTGCGGAAACGAAGT---------CCCACC
LJFgene1       AAAGGAGAAAGAAAAAACAAAAAAAAAATGTTGCGGAAACGAAGCGTCCCACCACCCACC
               ****** **  ** *** **** ********************          ******

LJFgene14      TTCTTTT---ATTCATCA---CATGATTCACATCTCATTTCTTATTTTCGGGTCACTTTG
LJFgene1       TTCTTTTTATATTCATCATTACATGATTCACATCTCATTCCATATTTTCGGGTCACATTC
               *******   ********   ****************** * ************** **

LJFgene14      TTAAATTAT---AAATAATTTCGTTCTGTGAAGGTACACACGTTCATAAGTTCCTTAATT
LJFgene1       TCAAATTATTATAACTAATTTCGTCATGTGAAGATAC----GTTCATAAGTTCCTTAATT
               * *******   ** *********  ******* ***    ******************

LJFgene14      TTCTCGAACCTTCATTTTCAGCTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTCC
LJFgene1       TTCTTGAACCTTTATTTTCAGCTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTCC
               **** ******* ************************************************

LJFgene14      TTCTGCAACCTCAAGTTTATCACCAAACCCAACAATGGTAGAAGAAGC---------TCT
LJFgene1       TTCTGCAACCTCAAGTTTATCACCAAACCCAACAACAATGGTAGAACCAATGCTTCTTCT
               ***********************************   * * **** *         ***

LJFgene14      CTTCGCCGTATTGTATTTTGTCAGAAGCATCACGATGACACACCCACCGACCAAATCAAC
LJFgene1       CTTCCCCGTATTGTATTCTGTCAGAAGCACAACGATGACACCCCCACCGACCAAATCAAC
               **** ************ ***********  ********** ******************

LJFgene14      CGAAGGTTCTTACTTCTTCACACTCACACTTTCTATTTCCTT---TCTATTGATTATTCG
LJFgene1       CGAAGGTTCTTACTTCTTCACACTCACACTCTCTCACTCCTTCTTTCTATTGATTATTTA
               ****************************** ***   *****   *************

LJFgene14      TAAC-----------CATCTTCTGAAATCTCGTTACATTTCAATTCTTTT-GTGTATTGAA
LJFgene1       TAACTATTCATACCCATCTTCTGAAATCTCTTTACATTTCAATTCTTTTTGTGTATTGAA
               ****          *************** ****************** **********

LJFgene14      GAGAACTCATATTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTCAACTTCGGGT
LJFgene1       GAGAACTCATATTGAGAAGCAGTGAAATAGCGACCATTGGTGCCATCTTCAACTTCGGGT
               ********************** ************************************

LJFgene14      ACCCCTCCTCTGTTTTTGCTCTGTTTTTTT---TTCTGGAAATTTTAGTTTTT-CATTTT
LJFgene1       ACCCCTCCTCTGTTTTTCCTCGGTTTTTTTCTTTTTTGGAAAATTTAGTTTTTTCATTTT
               ***************** *** ********   ** ****** ********** ******

LJFgene14      ATTTTGAATGTAAATTAAATTCGAGATTTGATTTTGTTAGTGGGTGTTGAGACCCTTTTG
LJFgene1       ATTTTGAATGTAAATTGTATTCAAGATTTGATTTTGTTGGTGGGTTTGGAGACCCTTTTG
```



```
LJFgene1        AAGCTTTTTCAAACATGCTCAAAGTTAAAGTGTTTCCATAATGT---TTTTTTGGGCCTT
                **** ******** * ********** **** ******* ****   ** ** *******

LJFgene14       AATTTTCGGTTAAGTAACTTGTGGTGTCAAAATTGCGTGTTGAGGAGGGTCTGGTAAACC
LJFgene1        AATTTTCG-TCAAGTAACTTGTGATGTCAAAATTGCGTGTTGAGGAGGGTCTGGTCAACT
                ******** * ************ ******************************** ***

LJFgene14       AGATTTAGGAGATAAGTAA----TCAGGGTTTATTTATTTATTGGATTTAGTCCAAACGG
LJFgene1        AGATTTAGGAGATAATTAAGCAGTCAGGGTTTATTTATTTATTGGATTTAATCCAAAAGG
                *************** ***    *************************** ****** **

LJFgene14       AACCTATGGCCATATTTGTGAGAATCACGTTGCAATAATAGACAATGGTGCACTTGGAAC
LJFgene1        AACCTATGAC-ATATTTGTGAGAATCACGTTGCAATA---GATAATGGTGCACTTGGAAC
                ******** * ************************   ** ******************

LJFgene14       AATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAGCCGAAATTGGACTATTGACTGTT
LJFgene1        AATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAACCGATATTGGACTATTGACTGTT
                ************************************ **** ******************

LJFgene14       GAACAATGTTGACTTGACAT--GAATTGGACTATTGGTCACACAC--ATTGGCCGGTGAA
LJFgene1        GAACATTGTTGACTTGACATAAGAAATGAACTATTGGTCACACACGCATTGGCCGGTGAA
                ***** **************  *** ** ****************  **************

LJFgene14       AGCAGTGCAATCTTAACTTTTTCTTTTTTTTTTGACAACTTTTTTTTTTTCTCTATCTTT
LJFgene1        AGTAGTGCAATCTTTACTTTTTCTTTTTTT------GAAATTTTTTTTTTCACTACCTTT
                ** ************ ***************       *  ************ *** ****

LJFgene14       GGTCCTTGTATTGAGCATGGTCCCACCAAAATCCAAACTTTATTATTGGACATAGATGAA
LJFgene1        GGTCCTTGTATTGAGCATGGTCCCACCAAAATCCAAACTTTATTATTGGACATAGATGAA
                ************************************************************

LJFgene14       TCATGATGTCACTTTGTTTTAATTTTTAGTTTCTTTTTCAGGAACTATAATCCTGAAGGA
LJFgene1        TCATGATGTCACTTTGTTTTAATATTTTCATTCTTTTTCAGGAACTATAATCCTGAAGGT
                ********************** ***   ****************************

LJFgene14       AGGAAAAAAACCTGTGAGCAGAGAGGAAGCAATGGAGGAACTGATAGACGTGGTAATAAAT
LJFgene1        AGGAAAAAAACCTGTGAGCAGGGAAGAAGCAATGGAGGAACTTATAGACGTGGTAATAAAT
                ******************** ** ***************** ******************

LJFgene14       CTAGCTGAACTGAAATCTTGAGTTATAACACTAGATTGCAATTTTCTTTTCCTTCCCTCA
LJFgene1        GCAACTGAACTGAAATCTTGAGTTAT--CACTGGATTGAAATTTTCTTTTCCTTCCCTCA
                  * **********************  **** ***** **********************

LJFgene14       TTTTATCAACATCGATTATAATTTATAAAATTTATAAAAGGAGTGATAGTAAATATACAC
LJFgene1        TTTTATCAACATTGATTATAATTTATAAAATTTATGAAAGGAGTGATAGTGAATATACAC
                ************ ********************** ************** *********

LJFgene14       TTTGTAACACACTATTCTTAACCCCACCCTTTTTATTATTGGTTAAAATTTATCAGAAAT
LJFgene1        TTTGTAACA--CTATTTCTAA---TACACTTTCTATTATCGGTTAAAATTTATTGAAAAC
                *********  *****  ***    ** **** ****** *************   ***

LJFgene14       TAGAAAGTCATG-AGTGGAACTCATTGAATAAAGAGTGAGACCTATG-TGATTTTAT---
LJFgene1        AGTGTTGTTAAATGGCGGCCATGGCGGCGCCATGGCGGAGTTGCGTAACGGTTTTCTGAA
                      ** *    * **   *    *    * *   ***     *   * **** *

LJFgene14       AATTTCTAATAAACTTATTATTA-ATAGTGTATTTAAAAGGATA-CATTGTGAACATTTC
LJFgene1        AAAACGCCACCGAATAACGGTGGCGTGGCGGATTAAAGATGGCGGCGCCATGGCGGCCGC
                **      *   * * *   *    * * * *** ** * *    *    **       *

LJFgene14       TCTAATT--------TATAATTGATTTGTAATAATAGCTGGTTAT--TTGCACTTTTCCA
LJFgene1        CATAGCCATGGCGGCCATGGCGGATGTGGCGGGGAGGCGGAAAATGGCAGAATTTTTTTT
                  **           **    *** **        ** *   **    * * ****

LJFgene14       ATCATT-GAAGTAAAGTTAAAACTAGTTCCGGATAATTTTACAGATCGAATCAACAACAC
LJFgene1        TTTGTCCGCGGTAGGAGTTGGGCTGACCCGATCCAACCCTAC--------CCGAAAACTT
                 *  *  *  ***    *    **    *     **   ***         * * ***

LJFgene14       CAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAGTACC
LJFgene1        AATGAAAACCATGACCCCCCCTACCTTTCAG--AACGCTGCTGCAACCCT-CGAAGCTTC
                 *   ***   * *** *   **  *   **  **     **  *  * *  * **   *
```

```
LJFgene14       AA---AGCTCAATCTTGGGGGTAAGTGTAACT--TACATCTAAG---GAAACTCATCTTG
LJFgene1        AACATAGCGCGACCTCGGAACCAGCCACGACCCCTGCAACCAGCCTCGGAACCAGCCGCG
                **   *** * * ** **    *      **   * ** * *     * ***    *  *

LJFgene14       AAGAAA--AATGATCATTTTATACATTTGAGATGATATCAAGAACCATCTCTAATTTCCT
LJFgene1        CTTGACCCGCTGCACACAGCAGCCAGCAGCGACGACCCATGGCGTGAACGGCGGCGACGA
                    *     **  **    *  **   * ** **      *    * *        *

LJFgene14       TCTCCCTTTTTTCTGTCATG-TGCTAACAGTTTGTGGATGATGTTGAGTTCTGGTTTCCA
LJFgene1        GCACGGCGTGAACAACGGCGACGCGAACGGA--GAAGACGACCCAGAACCGCGACCTCGA
                 * *    *   *      *  ** *** *   *  ** **    **     *   ** *

LJFgene14       CCCGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTCGGTTGGGAAAC-TTTGATTT
LJFgene1        CTTA-TACGGGTGGGTGGGCCAAAGCCTTTTTTTTGTTTTTGCTGCTTGACACCCCTTTT
                *    ** ****      *   *       * *   * *  * **    **      ***

LJFgene14       TGATGTGAACAGAAAAAGAATAAAGGTATGATTTCATAATTAATATGTGCTTTCTCT---
LJFgene1        TTATGTCTGCAGC-------TTTTTTTCCGTTTTTCTATTTGACACCCCCTTTTACTTTC
                * ****   ***        *     *  * ***  ** ** * *    ****  **

LJFgene14       -ATAGTTAGATAAAGAAATTCTTG----GTTCCAGGGTAAAACTCCCCTTCCT-TCATGT
LJFgene1        GACAGTCCCATTTTTAATTTTTTTTTCTATTTGACACCACAATTTTTTTTTCTGTTCAGT
                 * ***   **    ** ** **      **  *    * ** *    ** ** *   **

LJFgene14       CATGTC-----------AAACATTTTATACTTAAGTAGATTCACTAAATTTGAGT-CTCA
LJFgene1        CCTCTTTTAAATGGCTGAACCATCATCCTCTTTAATGAGTTTATTTGGTGTTGGTACTTG
                * * *            ** ***  *   *** * *   ** * *   * *  ** **

LJFgene14       AATGTTTTA--ACTTTATTCTAAATTAGTCACTTATTTTAACGGAAGG--TAAATTTGGT
LJFgene1        ATTATTGTATGAACTCATGAAACTTTTTAGTTTATTTGAATGCAATCCTTTGTTTTTTT
                * * ** **  *  * **   *  **  *   ****** ** * **    *   ***  *

LJFgene14       TGACTATGATGAG-----AAATACGTTGATATTTTTTAATTGGTAGAGATAAAG---AAT
LJFgene1        CAATTTCAATGAGTTTATATATATGTTTTTTTTTTTTTTTGGTCCGCCATGACTTCCGCC
                  * *   *****     * *** ***  * ******  * *   *  ** *

LJFgene14       ATTTTTTA-TGTACAATAAAGAGAGTATTTACTCCAGAGGATGCAAATCCCTTACTAAAT
LJFgene1        ATTTTCCGCTACGCCATCCGCCATATTTTTATGGCGGATTTTTTACTTTCCGCCATGAAC
                *****    *   * **        * ****   * **   *  *  * **    * **

LJFgene14       ATT--TTTGTGATGAAAAATCTTGGTTGCTGACA--GGCACTGAGACAAGAGTTGGA-GA
LJFgene1        CGCCATCCGCCATTAACAACATTGATTGAAAATTATGAAGTCATGAGAGGAGCTCATTGA
                     *  *  ** ** **  *** ***   *    *       ** * *** *    **

LJFgene14       AGAAAGGATGGACATCT-CAAGATACCATA-TGATTAATAAACTCAGGCTGA-ATTAGCA
LJFgene1        ATAAAGAGTGAGAACTTACATGATTTTGTAATTTCTAATAATTTTTTACTATTAATAAAG
                * ****  **   *  * ** ***    ** *   ******  *    **   * **

LJFgene14       TCAGCATCTAAGCAAATATTATTTCATATACTTTG--GACCTTGTATACTTTTGTATTAG
LJFgene1        AGTGTATTTAAATGGGTATGTTTTTGAACATTTTTCTAATTTCTAATCGATTTGTAATAA
                   * ** ***     ***  ***   * * ***    *  *   **   ****** **

LJFgene14       ATACAAATC----GCACAG----GATCATTGCAAGCAAACTTTTCTTAGATTTTTGGAAT
LJFgene1        TAGCTGGTTATTTGCACTTTCCCAATCATTG-AAGTAAAGTTAAACCAGTTCCGGATAAT
                   *   *     ****       ******* *** *** **     ** *      ***

LJFgene14       TGTAGAGAAATCATTGAGAACAGTACTTCAAACTCTCTCGGGGAAGG---AATGAAATGA
LJFgene1        TTTACAGATAGAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGA
                * ** *** *  **  * ****  *    **  *    *  ***  *   ** * ** **

LJFgene14       AGACCTTGCGC-CATATCTTCTTCAAGTTCATTAATTGGTCCACTTAT---TTACACCT-
LJFgene1        AGACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGGTAAGTGTAACTTACATCTA
                ****  *   *   *    *    *** *** *  * **   *  * *   ***** **

LJFgene14       -------TCACC--GAGTTCAATCTAATAATGTATGAATCTTG-----TTTCAAGAAATT
LJFgene1        AGGAAACTCATCATGAAGAAAAATTATCCTTTTATACATTTTAGATGATATCAAGATTCA
                       *** *  **    **  **    * ***  ** **      * ******

LJFgene14       TAATTTGTGTTTTGTTTCAAACG---TGTTCTTCATGATTCACT-----TTTGTTGT--A
LJFgene1        AGAACCATCTTTAATTTCCTTCTCCCTTTTTTCTGTCATGTGCTAACAGTTTGTGGATGA
                  *    * ***  ****   *    * ** *   * **   **     ***** *   *
```

```
LJFgene14       AGTTTGACTTTGCTTCTTCATGTATTAAAGCTTATCCTTGCG--ATCAATTTGGATGATG
LJFgene1        TGTTGAGTTCTGGTTTCCTCCGGGTAAGGGTTCTACTGTGGAGTATCGTTCTGCATCTCG
                 ***    * ** **      *  * *  * *   *  **    ***  * ** **   *

LJFgene14       CTTTAGG--CCTT---TCTGTCATCAAAATGTACGTACCTATGTCATGTTTTCAGAGGCT
LJFgene1        GTTGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATGATTCCATAATTC
                 **  *   * **   * **   * ** *   *   *   * *  *** ** ** *

LJFgene14       TTTTG-ATATCCTTGATGTCTTTACAAGAAAAGTC--GGTGC--ATATAGCATTTGGAAT
LJFgene1        ATATGTGCTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCAGGGTAAAACTCCCCTT
                 * **    * **  **   *  * **  *** **  *** *     **  * *     *

LJFgene14       CTGGTTATAATA---GGAGCATT----------GCGGAG--ACTAAGTATAGGGTGCA--
LJFgene1        TCCTTCATGTCATGTGAAGCATTTTTTACTCAAGTAGATCCACTAAATTTGAGTCTCAAA
                    * **   *   * ******          *  **   ***** * *  *   **

LJFgene14       --TTTTGACTTTGGTAACTCTGTTTC----TCAATTCACTTCCATACGTTGTTCAGAAAT
LJFgene1        TGTTTTAACTTTATTCTAAATGTTTTAACTTTATTTGAGTCTCAAATGTT--TTAACTGA
                  **** *****  *     *****     * * ** * *  ** * ***  * *

LJFgene14       TGTTCAATCTCATTGATCCTAACTAGAAA-ATTCGTCCAACAGC-AAAGAATGCAAAGCT
LJFgene1        AGGTAAATTTGGTTAACTATGATCAGAAATACATTAACAAGAGTTGAAGAATATTTTTTA
                 * * *** *  ** *   * *  ***** *      *** **   ******

LJFgene14       AGCAACTCAAGGACGAATTTTGTTGGCTTCTAGATAGTGGTTTGGCTTAAT---TTTGAG
LJFgene1        TGTACAATAAAGAGAGTATTTGCTCGAGAGAATGTAAATCCTTTTCTAAATATTTTTGTG
                 * *    ** **     **** * *     *  **     **  ** ***   **** *

LJFgene14       AATTTATTTGGTGGATTTTGG-ACGTA-TGAGGTTGAAGTTG----AGTTTTAATTGAAA
LJFgene1        ATGAAAAATAATGGTTGCTGGCAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGGCA
                *    *  *  *** *  *** * * * ****     *****    **     ** *  *

LJFgene14       TTTC---------CTTGTCCAACAAATGCAAGCAGATTTAGCATCATGA---AAAAAGAT
LJFgene1        TCTCAAGACACCATATGATGAAAAAACTTAGGCAGAATTCACATCAGCATCTAAGAAAAT
                * **           **   ** ***   * ***** **  *****  *   ** ** **

LJFgene14       A---------ATGCAATAAAAAAAAAAAAAGGTTAGAACAGGATAGAGAAGCACTCTAAA
LJFgene1        ATTGTTTCATATACATTGTAACCTTGTATACTTTTGTATTAGATACAAAATCTCACAAGA
                *         ** ** *  **      * *  ** * *   **** * ** * * * * *

LJFgene14       GCTATGAAAT----------------------------------------------------
LJFgene1        TCATTGAAAGCAAACTCTTCATGATTATTGGAATTGTAGAAATGATTGAGAACAGTACTT
                 *  *****

LJFgene14       ----------------------------------------------------------
LJFgene1        CAAACTCTCGGGGGAAGGAATGAAATGAAGATGTTACCCATATCTTTTTGAACTTCATTA


LJFgene14       ----------------------------------------------------------
LJFgene1        ATTGGTCCACTTATTTACTCTTTCGCTGAGTTCAATCTAACAATGTAGCATTCTTGTTTC


LJFgene14       ----------------------------------------------------------
LJFgene1        AAGAATTTTATGTTGTGTTGTTTTTCAAATGCATTTGCCATGATTCACCTTCGGTGCATA


LJFgene14       ----------------------------------------------------------
LJFgene1        TAGCATTTGGAATCTGATTATAATAGGAGCATTGCGGAGACCAAAAACAGGGTGCATT
```

LJFgene8 vs. LJFgene3

```
LJFgene8        ACCTCTAGTGTCCACTTAATGGTGAGATTTGAAATAATATGTATGAGATTCTAAGTTCAA
LJFgene3        ----------------------------------------------------------


LJFgene8        ATATTAAGTGTCATTGTTAAAGAAAAAACAGTATGATAT-GCTGGAATTTACTAAC-TAT
LJFgene3        ----------CCATAAAAAAGAAAAAAAAAAAGTCCCACCGCCCACCTTCTTTATCACAT
                          ***    **  ****** *   *   *  **     **   ** *  **

LJFgene8        AATCTTTATAATAAGTTTAGCATTTAGTGTATATTGACTTGAAAGATTCTTGTAGCAATT
```

```
LJFgene3       GATTCACATCTCATTCCTTATATTTGGTTCACATT--CTTAAATTATAAATATTTCGGTC
                **    **   *    *   **** **  * ***  *** **  **   * *  *  *

LJFgene8       TGCAGCAGTTTTATATAGATAG------TAACATTCTTAAAT-TACGTTTATAAGTTCCT
LJFgene3       TGT-GAAGATATATGTCCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCA
               **  * ** * *** *  ***       ***  ****  **  * * ***  *  * **

LJFgene8       TCATTTTCAGCTCCGAACAATGGCTTCTTCGTTCTCCTTCTGCACCCTCAAGTTTCGCAC
LJFgene3       ACAACAATGGCTT----CAATGGCATCTTCAAGCTCCTTCTGCAACCTCAAGTTCATCAC
                **      ***     ******* *****   *********** *********   ***

LJFgene8       CAAACCCAACGATAGTAGAAGCAGTGCTTCCTCTCTTCCCCGTATTCTATTCTGTCACAA
LJFgene3       CAAACCCAACAATGGTAGAAGAAG------CTCTCTTCCCCGTATTGTATTCTGTCAGAA
               ********** ** ******* **      **************** ********** **

LJFgene8       CCTCCACGATGACATTCACACACCCACTGACCAAATCAACCGAAGGTTCACTTCTCCACA
LJFgene3       GCACCACGATAG------CACACCCACCGACCAAATCAACCGAAGGTTCTTATTTCTTCA
                * *******        ********* **********************   * **  **

LJFgene8       CTCTCATGCACTTTCTCACAGCTTTCTAATTTCTATTGATTATTCATAATTATTCATGCA
LJFgene3       CACTC--GCACTTTCTAA-----TTCC--TTTCTATGGATTATTCATATCTATTC----A
               * ***  ********* *     ***   ******* ***********  *****    *

LJFgene8       TACCCATCTTCTGAAATCTCTTTACACGTCAATGATTTTGTATCTTAAAGACAACTCATA
LJFgene3       TACCCATCTTCTGAAATCTCTTTATATTTCAATTATTTTGTCTATTGAAGAGAACTCATA
               ************************ *  ***** ******* * ** **** ********

LJFgene8       TTGAGAAGCAGCGAAATAGCGACCATCGGTGCCATCTTCGACTTCAGGTACCCCCGGCCC
LJFgene3       TTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTGAACTTCGGGTACCCCT--CCT
               ************************** ***********  ***** ********   **

LJFgene8       CTCCTCTGTTTTTTAGAAATTTTT-CTTTTCATTTTATTTTGAACGTAAATTTAATTCAA
LJFgene3       CTGCT-TGTTTTTGGAAAATTTTTGTTTTTCATTTTATTTTGAATGTAAATTGAATTCAA
               ** ** *******   ********  ****************** ******* *******

LJFgene8       GATTTGATTTTGTTAGTGAGTAGTGAGACCCTTTTGGTTTTTAGTATTAGTTTTGTTTTG
LJFgene3       GATTTGATTTTGTTGGTGGGTTTGAAGACCCTTTTGGTTTTTAATTTCGGTTTTGTTTTG
               ************** *** **    ****************** * *  ***********

LJFgene8       TATTGAAAATGGGTAGTTATTAAAAAAGAGGGTATTGTGTGTTTTTTTTTTTTTTTTTTT
LJFgene3       TATTGGACATGGGTGGTGGTTAAAAAAGAGAAAATTGAGT-----TTGTGTCTTGTGTTT
               ***** * ****** **  ***********   **** **     ** * * ** * ***

LJFgene8       TTGTGGTGCAGTGGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCTTTA
LJFgene3       TGATGGTGCAGTGGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCATTA
               *  ************************************************** ***

LJFgene8       GCTCTGTGTCCGGTAACTAGGAACTGCGTATCAACCTCTGAGAATATCAGTGATCGCACT
LJFgene3       GCTCTGTGCCCGGCAACGAAGAATTGCGTGTCAACCTCTGAGAATATCAGTGATCGCACA
               ******** **** *** * *** ***** ****************************

LJFgene8       CATTATGCTCCTCTTTGGTAAAAACTTTCCTTGTTTTCCTCATTTTAATTTTAGCCTTCC
LJFgene3       CATTATGCTCCTCCATGGTAAAAG-TTTCCTTCTTTTTCTTATTTTAATTTTCACCTTGG
               *************  ********  ******* **** ** ***********  ****

LJFgene8       TTTTCAAACCGCCAAGTTAATTTTTTAACCGTGTATTGGTGGTTTTTTTTTTGGTCCTTGA
LJFgene3       ATTT------ATGGGATTA----------TATGTATTAAATGCATTTTTTTAATTGTGTG
                ***           ***           ******    *  *******  *  *

LJFgene8       ATGATACATAGGATTTT-TCAAATAATTGGTTCAAAGACTAATTA--TATTTATAATATG
LJFgene3       TTTGGACAAACTACTTAGTTAAGTGCTCATCTCATCATGTAAGTGCTTATGCATAAGTTG
                *   *** *  * **  * ** *  *    ***     *** *   ***  ****  **

LJFgene8       TGATTGTAGTTGGTCAAAGGGGAAATTGTTTGATTTTTATCATGTATAAAATGTTTTTGG
LJFgene3       TTTCTATAATAAA--AAAATAAAAATACATACGTATGAGTTGTTGTTGTAAGCTTTTTTC
               *   * ** *     ***    ****   *   * *   *  *   *  **  *****

LJFgene8       ACCAATAATAATAACAAATAAAATAAGATTAATTTTTAACTCAATTGGTTGATTAAGAGT
LJFgene3       TTAAGTTATTCTGG-AAATCTTATTGAAATAATCTGAAAACAACTTTTTTTTTTACATG-
                  * * **  *   ****   **   * **** *  **   * **  **  ***   *
```

```
LJFgene8       AAAAATATTTGTGATCTCTCATCAAAATAAAAAAATTTCTCTTCATTTTAAAACAAAAAT
LJFgene3       --ATCTGCTGAAAACAACTTATAGACATATGGTAATCACATATCATT---AAGTTAAGTT
                 *  *  *    *   ** **  * ***    ***  *   *****   **   **  *

LJFgene8       CACATCTTTAAAATTTAAAAGTCAGGTCCACACTTAAGATACTTAACTTAAACATAGGTA
LJFgene3       ---ATTCCAAACACTTACA---TAAACACTTATAAGAGAAAAACAA-TAAGAAATAAAAA
                  **    ** * *** *    *    *  *    *** *   ** * * * ***   *

LJFgene8       CTCAAACCAAAAATTTTCAAATATGTGAGATC---CCACACAAATTTTTATTTATTAGAA
LJFgene3       C---AAAATAAAATTTTCAATAAGTTAAAATTAGTTTATAAAAGTTTTTTTTTATTATAG
               *   **   ***********  *  * * **      * * ** ***** ******* *

LJFgene8       ATCTTAAAATAAATTAAATGTTTATCAGGAATTTGAATGTATTTTAGCCTATTTTCGAAC
LJFgene3       AAGCTCTCTTGTATTAGCCTCTCCTAAAGTTTTTTTTTCTTCATAATTTAGCTTAAAAAG
               *   *    *  ****     *  * * *  ***   * *   * *      **   **

LJFgene8       AAGGTTTTATCATACAAAATTTATGGATCCAACAAAATAAAAATAAGAAATTGTCATGGC
LJFgene3       AAACCTATTTCATTTTTTCATTTTATTTTCTTCTC--------TTGTAAATGCTTTTGGA
               **   * * ****       ** *   * *  *          *   ****  *  ***

LJFgene8       TTGGTTTCCAAAACTGAACCTCATGTTCAAGAACACGTGGAAGCACAACTGCAGCTTTGT
LJFgene3       GAAGTTTGTCCAAACATACCTT---TACACAAATACTTATAAG-ATAAGTCTAATTAAGC
                  ****    **    ****    * **  ** ** *  *** * ** *  *  *  *

LJFgene8       CTTAGCTCATCATCAAGCGCCAGTTAGCAATATTTTGCGGGAATAGACAATGAGACCAAA
LJFgene3       TTT--TTCA--AACACACTCAAAGTTAAAGTATTTCCATTTTGTTGTTTTT-----TTGG
                **   ***  * **  * * *  *   * *****        * *    *

LJFgene8       ACCTTCATCATGCACAAGTCTATTTTAGTTGCACTAGGGTTTTGGAGCCTTTCACAAGAC
LJFgene3       GCCTTAATTTTCGGTTAACTAACTT--GTGGTGTCTAAATTTCG---TCTTGAGGAGGGT
                **** **  *     *    * **  ** *        *** *    ***    * *

LJFgene8       CAAAACTAGATACGATTCATGCTAAAATAAAGCCTAAAGCCTTTTGGAATTATTCACAGC
LJFgene3       CTGA--TAAACCAGATTTAGGAGATAAGTAAGCAGTCAGTGTTT---ATTTATTTATTTT
               *  *  ** *   **** * *  * **  ****    **  ***   * ***** *

LJFgene8       AAAATCCGATGTTGACATCAACCATATGTAACATTATAATATATTACTAGCCTTTAAAAT
LJFgene3       TTGGATTTAGTCCAAAAGGAACC-TATGGCATATTTGTGAGAATCACG---TTGCAATAG
                       *     * *  **** ****  * ***       ** **     *  ** *

LJFgene8       TCAATTCCACATTCAACAAAATTTATTACCCTGTTCTCTCTACACTACGCAATTAGAAAA
LJFgene3       ACAATAGTGCACCTGGAACGATTTATCAC---GTTTTAAACAAGTAGTGCAACC-GATAT
                ****    **      *  ****** **   *** *    *      ****   ** *

LJFgene8       CAAAAGGTTGAGCGGAGAAAAATCTAGAAAGTGCTTAGAAATTCAACTAAATTTTGTCAA
LJFgene3       TGGACTATTGACTGTTGAACATTGTTGACTTGACATAAGAATTGGAC---AATTGGTCAC
                  *   ****  *  *** * * * **     * **  ****  **   * ** ****

LJFgene8       ACAGG-ACGAGCCAATAAATACGATTAATGACAGCTACGAGAAATGTATATTTTTAGTTT
LJFgene3       ACACACATTGGCCGGTGAA----AGTAGTGCAATCTTTACTTTTTATTTCTTTTTGACAA
               ***   *   ***  * **    * ** **  * **        * * * *****

LJFgene8       AGAAACACGTCTTTAGTTATA--TCATAACAAAAAATAAAAATTAAAGTTGTGTTTAAAT
LJFgene3       AAAAATTT-TCTCTACCTTTGGCCCTTGTTTGAGCATGGTCGCCACA---AAATCCAAAC
               * ***    *** **  * *    * *     *  **       * *      *  ***

LJFgene8       ATATGTTTTATTTAAAAAAATTAAAA----ATTAGGAAATTGTGTTTTAATTCATGACTT
LJFgene3       TTTATTATTGTATAGATGAATCATGATCTCACTTTGTTTTAGTATTTTCATTCTTTTTCA
                *   * ** * ** *  *** *  *    * *  *   * ** **** **** *

LJFgene8       TAAGTTAAAGTCACGAA--TTAAAAGATGGCTTTAATAAAAGTAAATTTAAAAAATAAAT
LJFgene3       GGAACTATAATCCTGAAGGTAGGAAAAAACCTGTGAACAGAGAGGAAGCAATGGAGGAAC
                  *  ** * **  ***  *   ** *   ** * *  * **   *   **   *  **

LJFgene8       AAAGAAGTCATGTTTGAAAATATACCTTTCGTTGAAATGTTGTGCTTT-----TAAAACA
LJFgene3       TGATAGA-CGTGGTAATAAATCTAACTGA-ACTGAAATTTTGAATTATATCACTGGATTG
                  * *   * ** *   **** ** **     ****** ***   * *     *  *

LJFgene8       CAACTTGCTCATTTATTGATTTTTTAAAAATAAATTTACCTATTTTGGGATTTTTTTTTT
LJFgene3       CAATTTTCTTTTCCTTCCCTCTTTTAA---TCAA--CATCGATTATAATTTATAAAATTT
               *** ** **  *   *   * ******   * **   * * *** *    * *    ***
```

```
LJFgene8       ATAAAAAATGCACCACTCTAGTAAGTAAACCTTTTTTTTATATATAAAAAATTGCATTCA
LJFgene3       ATAAAAGGGGTGATAAACT-GAAAGTAAATATACACTTTGTA-ATGCACTATT----CCT
               ******   *    *  ** * *******  *    *** ** **  *  ***     *

LJFgene8       GGTAAACAACATAATTAAGTGTTTACTGATTCATTGAAACACTTATGTATAAGTTGTTTA
LJFgene3       AACACACTTTCTA-TTATTCATTAA--AATTTATTGAAA-ATTACGAAGTCATGGGTGGA
                  * **    ** ***    ** *   *** ******* * *      * *   **  *

LJFgene8       TGTGATTGAAGAGAAAATAAAGTTAAATTATTTTCTTATAAATTGTAATATGTTTTCATG
LJFgene3       ATTCATTGAATAAAGAGTAAGACCTA--CATGATTTTGTAATTTCTAATAAACTCTTACT
                 * ****** * * * ***     *   **  * ** *** ** *****   * * *

LJFgene8       AGCTATGGAAAGTTTATTGAAATAAACTGAAAATAGATTGTGGATATTTCATAAATACAT
LJFgene3       ATTAATAAAGAATGTATTTAAA-------AGGGTATGTTGTCAACATTTCTCTAAT----
               *   **  * * * **** ***       *   **  ****  * *****   ***

LJFgene8       ATCTTAAAATTATTTTAAATATTCCAAACACTTATATAAATACTATAAGCACTTGTAATA
LJFgene3       ---TTATAGTTGATTTGTG-ATAATAGCTGGTTATTTG--CACTTTTCCTTCCAATCATT
                  *** * **  ***    **   *     **** *    *** *     *   * **

LJFgene8       GAAGAAAGATAAGAATGTAAAATAAATTGATTTTTTTTCCATAAGTTGATTTAAGTTAAA
LJFgene3       GAAGTAAA--------GTTAAACCAGTTCCGGACAATTTTGCAGATAGAATCAA--CAAC
               **** **         ** ***  * **        **    *  * ** * **   **

LJFgene8       ATCAACTTATGTATCATAACACCTCAGGTTTTGAAAAAGTTAAATGAGAGAGTTTTTAAC
LJFgene3       ACCAG---ACAAATTTT--CACCACGGATAGTTGAAAGG------AAAGAAGACTATATT
               * **    *   **  *  **** * * *  *  *** *       *   **  * **

LJFgene8       AAAGTTAAGTGTATAAGTTAAATTTAAGAGAAAAACACAATTTAT-TCTACCTTTTCTTT
LJFgene3       CGTGTGGAGTACCAAAGCTCAATTTTGGGGGTAAGTGTAACTTACATCTAAGGAAACTCA
                  **  ***    *** * *****  * *  **    ** ***  ****      **

LJFgene8       TCCCCATTTGTAATTGTTTATGGCCAATTTGATCCAAACAACATCTTTAGCCTAAATAAG
LJFgene3       TCACGAAGAAAAAT-GATAAT--TTTATACATTTGAGATGATATCA--AGATTCAAGAAC
               ** * *     *** * * **     **    *  * *  * ***   **  * ** **

LJFgene8       CTCTTCCAATCACACTCTAAG-TTTAAGTATGAT-TACTA-TGATGTTTAGATTTCATTG
LJFgene3       CATCTCTAATTTCCTTCCCTCCTTTTTCTGTCATGTGCTAACAGTTTGTAGATGATGTTG
               *   ** ***  *  **     ***   * * ** * ***    * * *****    ***

LJFgene8       TGTTGT--TTTTTATGACCTTAATTTTCCGTCGAGTAACGCCATATTTGAATTAAAGTTT
LJFgene3       AGTTCTGGTTCCCACCGGGTAAGGGTTCTACTGTGGAGTACCG-ATCTGCATCTCGGTTA
                *** *  **   *     * *   ***    * * *   **  ** ** **    ***

LJFgene8       GTAAATTTTAGATGAAATTCATTTTGCAAACTGAATTTGTAAAATCAACTCTTCCTTTAC
LJFgene3       GGAAACTTTGATTTTGA-------TGTGAACAGAAAAAGAATAA--AGGTGTGATTTCAT
               * *** ***   *   *       **  *** ***   * * **  *  * *   ** *

LJFgene8       TTCCAAAATCCAAACTTTATTATTGGACAAAGATGAATCGTGGAGTTACTTTGTTTTAAA
LJFgene3       AATTCATGTGTTTTCTCTATAGTTAGATAAAGA-AATTCTTGG---TTCCATGGTAAAAC
                    *  *     ** ***  ** ** *****  * ** ***   * *  ** *  **

LJFgene8       TTTTTAATTTTAATTATTGT-TCAGGAACTACAATCCTGAAGGTAGGAAAAACCCTGTGA
LJFgene3       TCCTCTTTCCTTCATGTCATGTCAAACATTTTATACTCAAGTAGATGATTCACTAAATTT
               *  *   *  *   * *  * ***   * *  *  *   *    * **   **    *

LJFgene8       GCAGAGAAGAGGCAATGGAGGAACTGATAGACGTGGTAATAAATCTAGCTGAAATCATAA
LJFgene3       GAGTCTCAAATGTTTTAACTTTATTC-TAAATTAGTCACTTATTTTAACTGAAGGTAAAT
               *      * * *   *      * *  ** *   *  * * * * ** *****   * *

LJFgene8       GTTATTTTC-ATGAATGCATTGCAATTTCCCTTTCCTAGCCTGTGTA--TCAACAAATGT
LJFgene3       TTGGTTAACTATGATCAGAAATACATTGACATTTTTTAATTGGTAGAGATAAAGAATATT
                *  **  * ****    *     ***  * ***  **    **  *  * ** **   *

LJFgene8       TATTATTTATAATAAGTTTAATTTTCATGCACTGACCGTATATAATAATT-TTATATTGA
LJFgene3       TTTTATGTACAATAAAGAGAGTATTTACTC-CAGAGGATGTAAATCCCTTGCTAAATATT
               * **** ** *****    * * ** *  * * **   * ** *    **  ** **

LJFgene8       TATCTAATCATAAATCATCATT--TAAATTATTTTAAGATAATTAATTTAAAAGTTAATA
LJFgene3       TTTGTGATGAAAAATCTTGGTTGTTGACAGGCACTGCGACAA--GAGTTGGAGAAGAAAG
```

```
                * * * ** * ***** *  **  * *       *  ** **   * **  *    **

LJFgene8        AATTTACCGTATATAATGAATTATAATTGAATAATTGTATAAAAAATTTATAATGTCACT
LJFgene3        GATGGGCATCTCAAGACACCATATGAT-GAATAAACTCAGGCAGAATTAACATCAGCATC
                 **   *     *  *     *** ** ******    *   * **** * *    **

LJFgene8        GTATAATTCTTTTTCTCATTTATAATTTGGTTGTAAGTTGTAGCTTGTAATAGCAGGTTA
LJFgene3        TAAGCAAATATTATTTCATATACTTTGTGACC-----TTGTATACATTTGTATTAGATA-
                  *  *    ** * **** **   * **        *****     *  **  ** *

LJFgene8        TTTACCTTTTTCCAATCATTGAAGTAAAGTTAAAATCCAGCTCTGGATAATTTTATAGAT
LJFgene3        -CAAATCTCACAGGATCATTGAAAGCAAACTTTTCTTTGATTATTGG-AATTGTAGAGAA
                   *   *      *********   **  *    *     * * *  **** ** ***

LJFgene8        AGAATCAACAACACCAGACAAATTTACACCACGAATAGTTGAAAGGAAGGAAGACTATAT
LJFgene3        ATCATTGAGAACAGTAC-----TTCAAACTCTCGGGGAAGGAATGAAATGAAGACCTTGC
                *  **  * ****  *      ** * **           *** * ** ******  *

LJFgene8        TCATGTGGAGTACCAAAGCTCAATCTTGGGGGTATGTGTAACTTACATCAAAAGGAAACT
LJFgene3        CCCATATCCTTCTTCAAGTTCATTAATTGGTCCGCTTAT---TTACTCTTTCACCAAGTT
                 *        *    *** *** *  * **      * *   ****      *  **  *

LJFgene8        CATCGTGGAGAAAAATAATAATTTTGTACATTTTAGATGATAATCAA-GAACCATCTCTA
LJFgene3        CA--ATCTAACAATGTATCGTTCTTGTT--TCAAGAATATTAATTTGTGTTTTGTTTCTA
                **   *  *  **  **    * ****   *     **  ****    *     * ****

LJFgene8        ATCCCTTCTCCCCTCCTTTTTATTTTTTCTGCCATGTGCTAGCAGTTT-GTGCAT-GATG
LJFgene3        ATGTGTTCTTCACGATTCACTTTTGTTGTAAGTTTGACTTAGCTTCTTCATGTATTAAAA
                **   **** * *   *   * ** **       **   ****   **  ** **  *

LJFgene8        TTGAGTTCTGGTTTCCACTGGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTCGGT
LJFgene3        CTTAACCTTGTGATCAATTTGGATACTTTAGGCCTTTCTCTTATCAAAATGTACCTATGT
                 * *    **   ** * * **  *   *    ** *    **** *  ** * **  **

LJFgene8        TGGGGAACTTTGATTTTGATGTGAATAAGAAAAGAATAAAGGTATGTTTGTATCATTCCT
LJFgene3        CATG--ATCTCGAAATTGTTCAATCTCATTGATCCTAACTAGAAAATTCGTCCAA-----
                   *  *  * **  *** *     * *   *     *   * *  ** **   *

LJFgene8        TTGTGCTGTCTCGGTAGTTAACATGAAGAAATGATTAAAAGATATTTTGTCCTTTAGGTT
LJFgene3        -------------------------------------------------------------


LJFgene8        TTTTGGTTATATTTAGTTTGATTTTTTATTTTTTAAAAGTTAAATTTAGTCCTTTATGTT
LJFgene3        -------------------------------------------------------------


LJFgene8        TTTAAAACGAATCAAAATGATCACTATTTCGAGATGTAAAACTTTTTTATTAGATTTACT
LJFgene3        -------------------------------------------------------------


LJFgene8        TACATGCAAATTATGCAACTTGACTAATGTTTGAAAAAAAAAATCATCAATATGATAAAA
LJFgene3        -------------------------------------------------------------


LJFgene8        ATAATGAGTGTCACAAGCAATTAAAAA
LJFgene3        ---------------------------
```

LJFgene8 vs. LJFgene14

```
LJFgene8         -------------------------------------------------------------
LJFgene14        TATATATACACACACACACGAATAAACAAATTTTTTTAAGAGTAAATTACATAACATCTT


LJFgene8         -------------------------------------------------------------
LJFgene14        GTGAGATTTTAAATTTTTTTATACATATTTAAAAAAAAAAGACTTACACAAATCTATCAA


LJFgene8         -------------------------------------------------------------
LJFgene14        TTAAATTTTAAAAAATTACACACGTCTCATAACTGTTTTTGAATAAATACTAACTAAAAT
```

```
LJFgene8         ------------------------------------------------------------
LJFgene14        TAAAAAAAATGTAGAAATGCATTATTATTTTTACCGAGTAAAAACATTCTTGATGCGCGA


LJFgene8         ----------------------------------------ACCTCTAGTGTCCACTTAA
LJFgene14        ATTTGACAAAAACCTTTTCGTACAGATAAGCATTTATGGATATTTTTAGTATCCA--AAA
                                                        *  * **** ****   **

LJFgene8         TGGTGAGATT-TGAAATAAT----ATGTATGAGATTCTAAGTTCAAATATTAAGTGTCAT
LJFgene14        TTGTCACTTTCTCAAACAATCGAAATATATACTATTTATTTTCTAAATATCTAAATCCAT
                 * ** *  ** * *** ***    ** ***   ***     *  ******  *    ***

LJFgene8         TGTTAAAGAAA-AAACAGTATGATA----TGCTG-------GAATTTACTAACTATAATC
LJFgene14        AAAAAGGAAAATAAATAAAAAAATAAAAATGTTGCGGAAACGAAGTCCCACCTTCTTTTA
                     *   *** *** *  *  ***    ** **       *** *  *    * *  *

LJFgene8         TTTATAATAAGTTTAGCATTTAGTGTATATTGACTTGAAAGATTCTTGTAGCAATTTGCA
LJFgene14        TTCATCACATGATTCACATCTCATTTCTTATTTTCGGGTCACTTTGTTAAATTATAAATA
                 ** ** * * * **  *** *  * * *  *     *     **  *  *   **    *

LJFgene8         GCA--GTTTTATATAGATAGTAACATTCTTAAATTAC---GTTTATAAGTTCCTTCATTT
LJFgene14        ATTTCGTTCTGTGAAGGTACACACGTTCATAAGTTCCTTAATTTTCTCGAACCTTCATTT
                      *** * *  ** **   ** *** *** ** *    ***    *  *********

LJFgene8         TCAGCTCCGAACAAT------------GGCTTCTTCGTTCTCCTTCTGCACCCTCAAGTT
LJFgene14        TCAGCTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCTCAAGTT
                 ******** ******            *** *****   *********** *********

LJFgene8         TCGCACCAAACCCAACGATAGTAGAAGCAGTGCTTCCTCTCTTCCCCGTATTCTATTCTG
LJFgene14        TATCACCAAACCCAACAATGGTAGAAGAAG------CTCTCTTCGCCGTATTGTATTTTG
                 *  ************* ** ******* **      ******** ******* **** **

LJFgene8         TCACAACCTCCACGATGACATTCACACACCCACTGACCAAATCAACCGAAGGTTC--ACT
LJFgene14        TCAGAAGCATCACGATGA------CACACCCACCGACCAAATCAACCGAAGGTTCTTACT
                 *** ** *  ********      ********* ********************  ***

LJFgene8         TCTCCACACTCTCATGCACTTTCTCACAGCTTTCTAATTTCTATTGATTATTCATAATTA
LJFgene14        TCTTCACACTC----ACACTTTCTA-----TTTCC--TTTCTATTGATTATTCGTAA---
                 *** *******     ********      ****   ***************** ***

LJFgene8         TTCATGCATACCCATCTTCTGAAATCTCTTTACACGTCAATGATTTTGTATCTTAAAGAC
LJFgene14        -----------CCATCTTCTGAAATCTCGTTACATTTCAATTCTTTTGTGTATTGAAGAG
                           ***************** *****  *****  ****** * ** ****

LJFgene8         AACTCATATTGAGAAGCAGCGAAATAGCGACCATCGGTGCCATCTTCGACTTCAGGTACC
LJFgene14        AACTCATATTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTCAACTTCGGGTACC
                 ********************************** ************ ***** ******

LJFgene8         CC----CGGCCCCTCCTCTGTTTTTT-----AGAAATTTTT-CTTTTCATTTTATTTTGA
LJFgene14        CCTCCTCTGTTTTTGCTCTGTTTTTTTTTCTGGAAATTTTAGTTTTTCATTTTATTTTGA
                 **    * *    * ***********      ********   *****************

LJFgene8         ACGTAAATTTAATTCAAGATTTGATTTTGTTAGTGAGTAGTGAGACCCTTTTGGTTTTTA
LJFgene14        ATGTAAATTAAATTCGAGATTTGATTTTGTTAGTGGGTGTTGAGACCCTTTTGGATTTTA
                 * ******* ***** ****************** **  ************** *****

LJFgene8         GTATTAGTTTTGTTTTGTATTGAAAATGGGTAGTTATTAAAAAAGAGGGTATTGTGTGTT
LJFgene14        GTTTGGGTTGTGTTTTGTATTGGAAATGGGTGGTTT----------GGGTTTTGTGT---
                 ** *  *** ************ ******** ***           **** ******

LJFgene8         TTTTTTTTTTTTTTTTTTTGTGGTGCAGTGGGAAAAAACCTGATTATCTTGGAGTGCAGA
LJFgene14        ---------------TTTGGTGGTGCAGTGGGAAAAAACCTGATTATCTTGGAGTGCAGA
                               *** ****************************************

LJFgene8         AAAACCCACCAGCTTTAGCTCTGTGTCCGGTAACTAGGAACTGCGTATCAACCTCTGAGA
LJFgene14        AAAACCCACCAGCATTAGCTCTGTGTCCGCCAACTAAGAACTGCGTGTCAACCTCTGAGA
                 ************* ***************  ***** ********* *************

LJFgene8         ATATCAGTGATCGCACTCATTATGCTCCTCTTTGGTAAAAACTTTCCTTGTTTTCCTCAT
LJFgene14        ATATCAGCGATCGCACACATTATGCTCCTCCATGGTAAAAG-TTCCCTTCTTTTTCTTAT
```

```
                ******* ******** *************  ********  ** **** **** ** **

LJFgene8         TTTAATTTTAGCCTTCCTTTTCAAACCGCCAAGTTAATTTTTTAACCGTGTATTGGTGGT
LJFgene14        TTTAATTTTCACCTTGGATTT-ATGGGATTATATGAATT----GAATGCAATTTTTTAAT
                 *********  ****   *** *       *  * ****     *  *    **  *  *

LJFgene8         TTTTTTTTGGTCCTTGAATGATACATAGGATTTTTCAAATAATTGGTTCAAAGACTAATT
LJFgene14        TGTGTTTGGATAAACAACTTAGTTA-ACTGCCCATCATGTAAGTG-CTTATGTATAAGGT
                 * * *** * *     * * *   * *       ***  *** **  * *   *  *  *

LJFgene8         ATATTTATAATATGTGATTGTAGTTGGTCAAAGGGGAAATTGTTTGATTTTTATCATGTA
LJFgene14        TTTTCTATAGTAAAAAAATGTA------CAAGTTGCAAGCTTTTTTCTTAAT-TTATTCT
                  * * **** **    * ****      ***   * **  * ***  **  * * **

LJFgene8         TAAAATGTTTTTGGACCAATAATAATAACAAATAAAATAAGATTAATTTTTAACT-CAAT
LJFgene14        TGAAATCTTATTGAAATAATC-TGAGAACAACTTT-------TTCTTTTTTACCTGTGAT
                 * **** ** *** *  ***  * * ***** *         **  ****** **   **

LJFgene8         TGGTTGATTAAGAGTAAAAATATTTGTGATCTCTCATCAAAATAAAAAAATTTCTCTTCA
LJFgene14        CTGAAAACAACTCATAGACATATCATAAACTTGGGATATAGATAAATTTTTTCATAAACA
                   *   *  *    ** * ****     *  *   **  * *****    **  *   **

LJFgene8         TTTTAAAACAAAAATCACATCTTTAAAATTTAAAAGTCAGGTCCACACTTAAGATACTTA
LJFgene14        CTTACAAGAGAAAAAAAAATACAAAAGAAAAATAAAATAAATTTTTCTATAAGCTAA--A
                  **  **   ****  * **    ** *   * **   *  *       **** **   *

LJFgene8         ACTTAAACATAGGTACTCAAACCAAAAATTTTCAAATATGTGAGATCCCACACAAATTTT
LJFgene14        ATTAGTGTATGTGTAAGCTAATTTGTAGTTCTT--TCATATTAGCTTCTCCAAAAGTTTT
                 * *     **  ***  * **     * ** *     ** * ** * *  ** ** ****

LJFgene8         TATTTATTAGAAATCTTAAAATAAATTAAATGTTTATCAGGAATTTGAATGTATTTTAGC
LJFgene14        TTTTTTTTTTAACTTATGCA-TAAGTTAAAT--TTAGCTTAAAGATAAATTTATTTCATT
                 * *** **  ** *  *  * *** ******  *** *   **  * *** ***** *

LJFgene8         CTATTTTCGAACAAGGTTTTATCATACAAAATTTATGGATCCAACAAAATAAAAATAAGA
LJFgene14        TTCTTCTCTTGTAAATGCTTTTGG---AGAA---ATGTATCCAAACAGACCTTTACAC-A
                  * ** **    **    ** *     * **   *** ******  * *     * *  *

LJFgene8         AATTGTCATGGCTTGGTTTCCAAAACTGAACCTCATGTTCAAGAACACGTGGAAGCACAA
LJFgene14        AGTACTAATAAGATAAGTCT---AATTAAGCCT----TTTCAAAACGC-TCAAAGTTCAA
                 * *  * **    *   *     ** * * ***    **  * *** * *  ***  ***

LJFgene8         CTGCAGCT-TTGTCTTAGCTCATCA----TCAAGCGCCAGTTAGCAATATTTTGCGGGAA
LJFgene14        GTATTTCCATTATGTTGATTCTTCGGGCCTTAATTTTCGGTTA--AGTAACTTGTGGTGT
                  *    *  ** * **   ** **     * **    * ****  * **  *** **

LJFgene8         TAGA-CAATGAGACCAAAACCTTCATCATGCACAAGTCTATTTTAGTTGCACTAGGGTTT
LJFgene14        CAAAATTGCGTGTTGAGGAGGGTCTGGTAAACCAGATTTAGGAGATAAGTAATCAGGGTT
                  * *     * *   *  *   **         **  * **    *   * * *  ** **

LJFgene8         TGGAGCCTTTCACAAGACCAAAACTAGATACGATTCATGCTAAAATAAAGCCTAAAGCCT
LJFgene14        T--ATTTATTTATTGGATTTAGTCCAAACGGAACCTATGGCCATATTT-GTGAGAATCAC
                 *  *    ** *   **   *  * * *    *   ***   * **   *    ** *

LJFgene8         TTTGGAATTATTCACAGCAAAATCCGATGTTGACATCAACCATATGTAACATTATAATAT
LJFgene14        GTTGCAATAATAGACAATGGTGCACT-TGGAACAATTTATCACGTTTTAAACCACGTGAA
                  *** *** **  ***        *  **     **  * **  * * * *  *    *

LJFgene8         ATTACTAGCCTTTAAAATTCAATTCCACATTCAACAAAATTTATTACCCTGTTCTCTCTA
LJFgene14        AGTAGT-GCAGCCGAAATTGGACTATTGACT-GTTGAACAATGTTGACTTGA----CATG
                 * ** * **     *****  * *    * *     **   * **  * **       *

LJFgene8         CACTACGCAATTAGAAAACAAAAGGTTGAGCGGAGAAAAATCTAGAAAGTGCTTAGAAAT
LJFgene14        AATTGGACTATTGGTCAC--ACACATTGGCCGGTGAAAGCAGTGCAAT---CTTAACTTT
                  * *   * *** *  *   * *  ***  *** ****    *  **    ****    *

LJFgene8         TCAACTAAATTTTGTCAAACAGGACGAGCCAATAAATACGATTAATGACAGCTACGAGAA
LJFgene14        TTCTTTTTTTTTTGACAA----------CTTTTTTTTTTTCTCTATCTTTGGTCCTTGTA
                 *    *   ***** ***          *   *   *    *  **    * * *  * *

LJFgene8         ATGTATATTTTTAGTTTAGAAACACGTCTTTAGTTATATCATAACAAAAAATAAAAATTA
```

```
LJFgene14       TTGAGCATGGTCCCACCAAAATCCAAACTTTA-TTATTGGACATAGATGAATCATGATGT
                 **   **  *      * ** *    ***** ****   * *   *  *** *  **

LJFgene8        AAGTTGTGTTTAAATATATGTTTTATTTAAAAAAATTAAAAATTAGGAAATTGTGTTTTA
LJFgene14       CACTT-TGTTTTAATTTTTAGTTTCTTTTTCAGGAACTATAATCCTGAAGGAAGGAAAAA
                 * ** ***** *** * *  *** ***   *  *   * ***   ***     *    *

LJFgene8        ATTCATGACTTTAAGTTAAAGTCACGAATTAAAAGATGGCTTTAATAAAAGTAAATTTAA
LJFgene14       ACCTGTGAGCA-GAGAGGAAGCAATGGAGGAACTGATAGACGTGGTAA---TAAATCTAG
                *    ***     **   ***  * * *  **  *** *   *  ***   ***** **

LJFgene8        --AAAATAAATAAAGAAGTCATGTT-TGAAAATATACCTTTCGTTGAAATGTTGTGCTTT
LJFgene14       CTGAACTGAAATCTTGAGTTATAACACTAGATTGCAATTTTCTTTTCCTTCCCTCATTTT
                   ** * **      *** **      * * *  *  **** **    *       ***

LJFgene8        TAAAACACAACTTGCTCATTTATTGATTTTTTAAAAATAAATTTACCTATTTTGGGATTT
LJFgene14       ATCAACATCGATTA-TAATTTATAAAATTTATAAAAGGAGTGATAGTAAATATACACTTT
                   ****    **  * ******  * *** *****  *    **   * * *    ***

LJFgene8        TTTTTTTATAAAAAATGCACCACTCTAGTAAGTAAACCTTTTTTTTTATATATAAAAAATT
LJFgene14       GT-----------AACACACTATTCTTA-ACCCCACCCTTTTTATTATTGGTTAAAATTT
                 *           **  *** * ***   *    * ******* ****   * **** **

LJFgene8        GCATTCAGGTAAACAACATAATTAAGTGTTTA-CTGATTCATTGAAACACTTATGTATAA
LJFgene14       A---TCAG--AAATTAGAAAGTCATGAGTGGAACTCATTGAATAAAG-AGTGAGACCTAT
                    ****  ***  * * * * * * **  * ** *** * * **  * * *    **

LJFgene8        GTTGTTTATGTGATTGAAGAGAAAATAAAGTTAAATTATTTTCTTATAAATTGTAATATG
LJFgene14       GTGATTT-TATAATTTCTAATAAACTTATTATTAATAGTGTATTTAAAAG----GATACA
                **  *** * * ***    * *** * *   * ***  * *  *** **      ***

LJFgene8        TTTTCATGAGCTATGGAAAGTTTA-TTGAAATAAACTGAAAATAGATTGTGGATATTTCA
LJFgene14       TTGTGAACATTTCTCTAATTTATAATTGATTTGTAATAATAGCTGGTTATTTGCACTTTT
                ** * *  *  * *  **  * ** **** *  * * * *   * ** *    * **

LJFgene8        TAAATACATATCTTAAAATTATTTTAAATATTCCAAACACTTATATAAAT---ACTATAA
LJFgene14       CCAATCATTGAAGTAAAGTTAAAACTAGT--TCCGGATAATTTTACAGATCGAATCAACA
                  ***   *    **** ***     * *  ***  * * ** ** * **   *  *  *

LJFgene8        GCACTTGTAATAGAAGAAAGATAAGAATGTAAAATAAATTGATTTTTTTTCCATAAGTTG
LJFgene14       ACACCAGACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAG
                 ***  *  * *     *  *     *  * ***  **   *   * * * * *  *  *

LJFgene8        ATTTAAGTTAAAATCAACTTATGTATCATAACACCTCAGGTTTTGAAAAAGTTAAATGAG
LJFgene14       TACCAAAGCTCAAT---CTTGGGGGTAAGTGTAACTTACATCTAAGGAAACTCATCTTGA
                    **     ***   ***  *  * *    * ** *  * *    *** * *  *

LJFgene8        AGAGTTTTTAACAAAGTTAAGTGTATAAGTTAAATTTAAGAGAAAAACACAATTTATTCT
LJFgene14       AGAAAAATGATCATT-TTATACATTTGAGATGATATCAAGAACCATCTCTAATTTCCTTC
                ***    * * **   ***    * * ** * *  * ****   *     *****  *

LJFgene8        ACCTTTTCTTTTCCCCATTTGTAATTGTTTATGGCCAATTTGATCCAAACAACATCTTTA
LJFgene14       TCCCTTTTTTCTGTCATGTGCTAACAGTTTGTGGATGAT---GTTGAGTTCTGGTTTCCA
                 ** *** ** *  *   *  ***  **** ***   **    *  *       * *  *

LJFgene8        GCCTAAATAAGCTCTTCCAATCACACTCTAAGTTTAAGTATGATTACTATGATGTTTAGA
LJFgene14       CCCGGTAAGGGTTCTACTG--TGGAGTATCGATCT--GCATCTCGGTTGGGAAACTTTGA
                 **   *   * *** *       * * *   * *  * **      *  **   ** **

LJFgene8        TTTCATTGTGTTGTTTTTTATGACCTTAATTTTCCGTCGAGTAACGCCATATTTGAATTA
LJFgene14       TTTTGATGTGAA------CAGAAAAAGAAT----AAAGGTATGATTTCATAATTAATATG
                ***   ****         *  *    ***        *  * *   **** ** *  *

LJFgene8        AAGTTTGTAAATT-TTAGATGAAATTCATTTTGCAAACTGAATTTGTAAAATCAACTCTT
LJFgene14       TGCTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCAGGGT-----AAAACTCCCCTT
                   *** *  **  ****** **     * ***    * *  *     *** *  * ***

LJFgene8        CCTTTACTTCCAAAATCCAAACTTTATTATTGGACAAAGATGAATCGTGGAGTTACTTTG
LJFgene14       CCTTCATGTC--ATGTCAAACATTTTATACTT----AAGTAGATTCACTAAATT--TGAG
                **** *  **  *  ** **  ***  ** *     ***  ** **    * **  *  *
```

```
LJFgene8         TTTTAAATTTTTAATTTTAATTATTGTTCAGGAACTACAATCCTGAAGGTAGGAAAAACC
LJFgene14        TCTCAAATGTTTTAACTTTATTCTAAATTAGTCACT--TATTTTAACGGAAGGTAAATTT
                 * * **** *** *  ** *** *   * **  ***   **  * * ** *** ***

LJFgene8         ------CTGTGAGCAGAGAAGAGGCAATGGAGG--AACTGATAGACGTGGTAATAAATCT
LJFgene14        GGTTGACTATGATGAGAAATACGTTGATATTTTTTAATTGGTAGAGATAAAGAATATTTT
                       ** ***  *** *   *   **       ** ** ****  *    *  * * *

LJFgene8         --AGCTGAAATCATAAGT-TATTTTCATGAATGCATTGCAATTTCCCTTTCCTAGCCTGT
LJFgene14        TTATGTACAATAAAGAGAGTATTTACTCCAGAGGATGCAAATCCCTTACTAAATATTTTT
                   *  *  *** *  **  ***** *   *  * **   ***  *    *       * *

LJFgene8         GTATCAACAAATGTTA-TTATTTATAATAAGTTTAATTTTCATGCACTGACCGTATATAA
LJFgene14        GTGATGAAAAATCTTGGTTGCTGACAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGG
                 **    * **** **  **  * * *   * *   *      *      *      **

LJFgene8         TAATTTTATATTGATATCTAATCATAAATCATCATTTAAATTATTTTAAGATAATTAATT
LJFgene14        ACATCTCAAGATACCATATGATTAATAAACTCAGGCTGAATTAGCATCAGC--ATCTAAG
                   ** * *   *   ** * ** *  ** *      * *****   * **   **  *

LJFgene8         TAAAAGTTAATAAATTTACCGTATATAATGAATTATAATTGAATAATTGTATAAAAAATT
LJFgene14        CAAATATTATTTCATATACTTTGGACCTTG---TATACTTTTGTA-TTAGATACAAATCG
                  ***  *** *  ** ***  *  *   **   **** **   ** **  *** ***

LJFgene8         TATAATGTCACTGTA--TAATTCTTTTTCTCATTTATAATTTGGTTGTAAGTTGTAGCTT
LJFgene14        CACAGGATCATTGCAAGCAAACTTTTCTTAGATTTTTGGAATTGTAGAGAA--ATCATTG
                  * *   *** ** *   **   *** *   **** *    * ** *  *    *   *

LJFgene8         GTAATAGCAGGTTATTTACCTTTTTCCAATCATTGAAGTAAAG--TTAAAATCCAGCTCT
LJFgene14        AGAACAGTACTTCAAACTCTCTCGGGGAAGGAATGAAATGAAGACCTTGCGCCATATCTT
                   ** ** *  * *    *  *     **  * **** * ***   *     *      *

LJFgene8         GGATAATTTTATAGATAG----AATCAACAACACCA--GACAAATTTACACCACGAATAG
LJFgene14        CTTCAAGTTCATTAATTGGTCCACTTATTTACACCTTCACCGAGTTCAATCTAATAATGT
                     ** ** **  ** *    * * *   *****     * * ** *  * *  ***

LJFgene8         TTGAAAGGAA----GGAAGACTATATTCATGTGGAGTACCAAAGCTCAATCTTGGGGGT-
LJFgene14        ATGAATCTTGTTTCAAGAAATTTAATTTGTGTTTTGTTTCAAA-CGTGTTCTTCATGATT
                  ****            * * *  ***  ***   **  **** *    ****   * *

LJFgene8         -ATGTGTAACTTACATCAAAAGGAAACTCATCGTGGAGAAAAA-TAATAATTTTGTACAT
LJFgene14        CACTTTTGTTGTAAGTTTGACTTTGCTTCTTCATGTATTAAAGCTTATCCTTGCGATCAA
                  *  * *    **  *   *       ** ** ** *  ***  * **  **  *  **

LJFgene8         TTTAGATGATAATCAAGAACCATCTCTAATC----CCTTCTCCCCTCCTTTTTATTTTT-
LJFgene14        TTTGGATGATGCTTTAGGCCTTTCTGTCATCAAAATGTACGTACCTATGTCATGTTTTCA
                 *** ******  *  **  *  *** * ***      * *   ***   *  * ****

LJFgene8         -----TCTGCCATGTGCTAGCAGTTTGTGCATGA------------TGTTGAGTTCTGGT
LJFgene14        GAGGCTTTTTGATATCCTTGATGTCTTTACAAGAAAAGTCGGTGCATATAGCATTTGGAA
                      * *   ** * ** *  ** * * ** **            * * *  **  *

LJFgene8         TTCCACTG-GGTAAGGGTTCTACTGTGG---AGTATCGATCTGCATCTCGG-TTGGGGAA
LJFgene14        TCTGGTTATAATAGGAGCATTGCGGAGACTAAGTATAGGG-TGCATTTTGACTTTGGTAA
                 *     *    ** * *   * * * *    ***** *   ***** * *  ** ** **

LJFgene8         CTTTGATTTTGATGTGAATAAGAAAAGA-ATAAAGGTATGTTTGTATCATTCCTTTGTGC
LJFgene14        CTCTGTTTCTCAATTCACTTCCATACGTTGTTCAGAAATTGTTCAATC--TCATTGATCC
                 ** ** ** * *  * * *   * * *   *  **  **  **  ***  ** **  * *

LJFgene8         TGTCTCGGTAGTTAA--CATGAAGAAATGATTAAAAGATATTTTGTCCTTTAGGTTTTTT
LJFgene14        TAACTAGAAAATTCGTCCAACAGCAAAGAATGCAAAGCTAGCAACTCAAGGACGAATTTT
                 *  ** *  * **    **  *  ***  **  **** **     **    * *  ****

LJFgene8         GGTTATATTTAGTTTGATTTTTTATTTTTTAAAAGTTAAATTTAGTCCTTTATGTTTTTA
LJFgene14        GTTGGCTTCTAGATAGTGGTTTGGCT---TAATTTTGAGAATT--TATTTGGTGGATTTT
                 * *    * *** * *   ***   *   ***   * * * **  *  **  **  ***

LJFgene8         AAACGAATCAAAATGATCACTATTTCGAGATGTAAAACTTTTTTATTAGATTTACTTACA
LJFgene14        GGACGTATGAGGTTGAAGTTGAGTTTTAATTG--AAATTTCCTTGTCCAACAAA--TGCA
                   *** ** *   ***     * **  *  **  *** **  ** *   *   *  * **
```

```
LJFgene8         TGCAAATTATGCAACTTGACTAATGTT--TGAAAAAAAAAATCATCAATATGATAAAAAT
LJFgene14        AGCAGATTTAGCATCATGAAAAAAGATAATGCAATAAAAAAAAAAAAAAGGTTAGAACAGG
                  *** ***  *** * ***  ** * *  ** ** ******  *  **  * * ** *

LJFgene8         AATGAGTGTCAC----AAGCAATTAAAAA
LJFgene14        ATAGAGAAGCACTCTAAAGCTATGAAAT-
                 *  ***   ***    **** ** ***
```

LJFgene8 vs. LJFgene1

```
LJFgene8        ACCTCTAGTGTCCACTTAATGGTGAGATTTGAAATAATATGTATGAGATTCTAAGTTCAA
LJFgene1        -------------------------------------TAAAAAGGAGAAAGAAAAAACAA
                                                     **   * ****    **   ***

LJFgene8        ATATTAAGTGTCATTGTTAAAGAAAAAACAGTATGATATGCTGGAATTTACTAACTATAA
LJFgene1        AAAAAAAATGTTGCGG----AAACGAAGCGTCCCACCACCC----ACCTTCTTTTTATAT
                * *  ** ***    *    * *  ** *        *  *    *  * **   ****

LJFgene8        TCTTTATAATAAGTTTAGCATTT-AGTGTATATTGACTTGAAAGATTCTTGTAGCAATTT
LJFgene1        TCATCATTACATGATTCACATCTCATTCCATATTTTCGGGTCACATTCTCA-AATTATTA
                ** * ** * * * **  *** * * *  *****  *  *  * *****   *   ***

LJFgene8        GCAGCAGTTTTAT-ATAGATAGTAACATTCTTAAATTAC---GTTTATAAGTTCCTTCAT
LJFgene1        TAACTAATTTCGTCATGTGAAGATACGTTCATAAGTTCCTTAATTTTCTTGAACCTTTAT
                  *  * ***  * **    **  ** *** *** ** *    ***    *  **** **

LJFgene8        TTTCAGCTCCGAACAAT------------GGCTTCTTCGTTCTCCTTCTGCACCCTCAAG
LJFgene1        TTTCAGCTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTCCTTCTGCAACCTCAAG
                ********** ******            *** *****   *********** *******

LJFgene8        TTTCGCACCAAACCCAACGATA---GTAGAAGCAGTGCTTCCTCTCTTCCCCGTATTCTA
LJFgene1        TTTATCACCAAACCCAACAACAATGGTAGAACCAATGCTTCTTCTCTTCCCCGTATTGTA
                ***  ************* * *   ****** ** ****** *************** **

LJFgene8        TTCTGTCACAACCTCCACGATGACATTCACACACCCACTGACCAAATCAACCGAAGGTTC
LJFgene1        TTCTGTCAGAAGCACAACGATGA------CACCCCCACCGACCAAATCAACCGAAGGTTC
                ******** ** * * *******      *** ***** **********************

LJFgene8        --ACTTCTCCACACTCTCATGCACTTTCTCACAGCTTTCTAATTTCTATTGATTATTCAT
LJFgene1        TTACTTCTTCACACTC----ACACTCTCTCAC----TCCTTCTTTCTATTGATTATTTAT
                  ****** *******     **** ******    * **  *************** **

LJFgene8        AATTATTCATGCATACCCATCTTCTGAAATCTCTTTACACGTCAATGATTTT-GTATCTT
LJFgene1        AACTATTC----ATACCCATCTTCTGAAATCTCTTTACATTTCAATTCTTTTTGTGTATT
                ** *****    **************************  *****  **** ** * **

LJFgene8        AAAGACAACTCATATTGAGAAGCAGCGAAATAGCGACCATCGGTGCCATCTTCGACTTCA
LJFgene01        GAAGAGAACTCATATTGAGAAGCAGTGAAATAGCGACCATTGGTGCCATCTTCAACTTCG
                 **** ****************** ************** ************ *****

LJFgene8        GGTACCCC----CGGCCCCTCCTCTGTTTTTT--------AGAAATTTT--TCTTTTCAT
LJFgene1        GGTACCCCTCCTCTGTTTTTCCTCGGTTTTTTTCTTTTTTGGAAAATTTAGTTTTTTCAT
                ********    * *    ***** *******         **** ***  * *******

LJFgene8        TTTATTTTGAACGTAAATTTAATTCAAGATTTGATTTTGTTAGTGAGTAGTGAGACCCTT
LJFgene1        TTTATTTTGAATGTAAATTGTATTCAAGATTTGATTTTGTTGGTGGGTTTGGAGACCCTT
                *********** *******  ******************** *** **   *********

LJFgene8        TTGGTTTTTAGTATTAGTTTTGTTTTGTATTGAAAATGGGTAGTTATTAAAAAAGAGGGT
LJFgene1        TTGGATTTTAGTTTCAGTTTTGTATTGTATTGGAAATGGGTGGTGGTTAAAAAAGAGAAA
                **** ******* * ******** ******** ******** **  ***********

LJFgene8        ATTGTGTGTTTTTTTTTTTTTTTTTTTTTGTGGTGCAGTGGGAAAAAACCTGATTATCTT
LJFgene1        ATTGAGT-----TTGGGTTTTGTGTTTTGGTGGTGCAGTGGGAAAAAACCTGATTATCTT
                **** **     **   **** * **** ******************************

LJFgene8        GGAGTGCAGAAAAACCCACCAGCTTTAGCTCTGTGTCCGGTAACTAGGAACTGCGTATCA
LJFgene1        GGAGTGCAGAAAAACCCACCAGCATTAGCTCTGTGTCCGGCAACTAAGAACTGCGTGTCA
                ********************** ********** ***** ********* ***

LJFgene8        ACCTCTGAGAATATCAGTGATCGCACTCATTATGCTCCTCTTTGGTAAAAACTTTCCTTG
```

```
LJFgene1       ACCTCTGAGAATATCAGTGATCGCACACATTATGCTCCTCCATGGTTAAAG-TTCCCCTC
               ************************** *************  **** ***  ** ** *

LJFgene8       TTTTCCTCATTTTAATTTTAGCCTTCCTTTTCAAACCGCCAAGTTAATTTTTTAACCGTG
LJFgene1       TTTTTCTTATTTTAATTTTCACCTTCGATTT-ATGGGATTATATGAATTAAATGCAATTT
               **** ** ***********  *****  *** *       *  * ****   *     *

LJFgene8       TATTGGTGGTTTTTTTTTGGTCCTTGAATGATACATAGGATTTTTCAAATAATTGGTTCA
LJFgene1       TTTTTACTTGTCTGTTTGGATAAACAACTTAGTTA-AGTATTCATCATGTAAGTG-CTTA
               * **      * * *** * *     * * *   * ** ***  ***  *** **  * *

LJFgene8       AAGACTAATTATATTTATAATATGTGATTGTAGTTGGTCAAAGGGGAA--ATTGTTTGAT
LJFgene1       TGTATAAGTTGTTTCTATAATAAATAAAAATGAGGAGGGAAAGTTATTCCAAAAATCACT
                  *  * ** * * *******  * *   *     *  ****       *    *   *

LJFgene8       TTTTATCATGTATAAAATGTTTTTGGACCAATAATAATAACAAA---TAA--AATAAGAT
LJFgene1       TTAAAAGAGGTACTCACTTTATTTTAATTATTGATTTTTTTAAAATTTAATGATTAAGAT
               **  *  * ***   * * * ***  *  * * * **  *   ***   ***  * ******

LJFgene8       TAATTTTTAACT-CAATTGGTTGATT--AAGAGTAAAAATATTTGTGATCTCTCATCAAA
LJFgene1       TAATTATTTATTATAATTATTAGATTCTAAAAAAATAAATAATAAGGATAATAAATAACT
               ***** ** * *  ****  * ****  ** *  * ***** *   ***     ** *

LJFgene8       ATAAAAAAATTTCTCTTCATTTTAAAACAAAAATCACATCTTTAAAATTTAAAAGTCAGG
LJFgene1       CTAAAAAAATTATTCTT-AGAGCAAGTTGATGTTTTCCTAAAAAAAATATATAAATTGTT
                **********  **** *    **    *   *  * *    ***** ** ** *

LJFgene8       TCCACACTTAAGATACTTAACTTAAACATAGGTACTCAAACCAAAAATTTTCAAATATGT
LJFgene1       TATATAAGTCGTAAGCTT--TTCGGAAATTATTCTTGAAATC----TTATTGAAATA---
               *  * *  *   *  ***   *   * **   *  * *** *     * ** *****

LJFgene8       GAGATCCCACACAAATTTTTATTTATTAGAAATCTTAAAATAAATTAAATGTTTATCAGG
LJFgene1       ---ATCTGAAAACAACTTTTTTTTACATGA--TTTGAAAACAACTTATAGACATATCATA
                  ***  * *  ** **** ****   **  * * **** ** *** *    *****

LJFgene8       AATTTGAATGTATTTTAGCCTATTTTCGAACAAGGTTTTATCATACA----AAATTTATG
LJFgene1       A--TCACATATCATTAAGTTATTTTATTAAGTCATTTTCATAATTTATTTCAAACACTTA
               *  *   ** *  ** **    ***   **     *** ** **  *    ***    *

LJFgene8       GATCCA-ACAAAATAAAAATAAGAAATTGTCATGGCTTGGTTTCCAAAACTGAACCTCAT
LJFgene1       CATAAATACTTATAAGAGAAAATAAAATAAAATTAATTTCTTTATAAGCTATAAAATTA-
                **  * **  *  * * * ** *** *   **   **  ***  **     **  * *

LJFgene8       GTTCAAGAACACGTGGAAGCACAACTGCAGCTTTGTCTTAGCTCATCATCAAGC-GCCAG
LJFgene1       GTTAATGTATAAGTCAATAGGTAGAAGCTCTCTCGTATTAACTCTTCAAAAGTTTATTTT
               *** * * * * **  *     *   **    * ** *** *** ***  *

LJFgene8       TTAGCA-ATATTTTGCGGGAATAGACAATGAGACCAAAACCTTCATCATGCACAAGTCTA
LJFgene1       TTAACTTATATATAAGCTAAATTTAACTTAAAAGAGAAACTTATTTCATTTTTTCTCTTC
               *** *  **** *      ***  *   * * *   **** *   ****         *

LJFgene8       TTTTAGTTGC-ACTAG--GGTTTTGGAGCCTTTCACAAGACCAAAACTAGATACGATTCA
LJFgene1       CTTTCTTTTCTAGTAAATGTTTTTATAAAAGTTTACCCAAACAGAACTTTACACAAGTAC
                ***  ** * * **   * ****  *    ** **   * ** ****  * ** * *

LJFgene8       TGCTAAAATAAAGCCTA--AAGCCTTTTGGAAT-TATTCACAGCAAAATCCGATGTTGAC
LJFgene1       TTATAAGATAAGTCTAATTAAGCTTTTTCAAACATGCTCAAAGTTAAAGT--GTTTCCAT
               *  *** ****  *  *  **** ****  **  *  *** **  ***     * *  *

LJFgene8       ATCAACCATATGTAACATTATAATATATTACTAGCCTTTAAAATTCAA-TTCCACATTCA
LJFgene1       AATGTTTTTTTGGGCCTTAATTTTCGTCAAGTAACTTGTGATGTCAAAATTGCGTGTTGA
               *       * **   * * **  *     * ** * * * *  *  ** ** *   ** *

LJFgene8       ACAAAATTTATTACCCTGTTCTCTCTACAC---TACGCAATTAGAAAACAAAAGGTTGAG
LJFgene1       GGAGGGTCTGGTCAACTAGATTTAGGAGATAATTAAGCAGTCAGGGTTTATTTATTTA--
                 *   * *  *   **    *    * *    ** *** * **     *     **

LJFgene8       CGGAGAAAAATCTAGAAAGTGCTTAGAA---ATTCAACTAAATT---TTGTCAAACAGGA
LJFgene1       TTGGATTTAATCCAAAAGGAACCTATGACATATTTGTGAGAATCACGTTGCAATAGATAA
                 *     **** * ** *  * **  *   ***      ***    ***  * * *  *
```

```
LJFgene8       CGAGCCAATAAATACGATTAATGAC------AGCTACGAGAAATGTATATTTTTAGTTTA
LJFgene1       TGGTGCACTTGGAACAATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAACCGATATT
                *   ** *    ** *** ** **      * * *** ****    *        * *

LJFgene8       GAAACACGTCTTTAGTTATATCAT-AACAAAAAATAAAAATTAAAGTTGTGTTTAAATAT
LJFgene1       GGACTATTGACTGTTGAACATTGTTGACTTGACATAAGAAATGAACTATTGGTCACACAC
               * *  *     *     * **  *  **   * **** ** * ** *  ** * * * *

LJFgene8       ATGTT--TTATTTAAAAAAATTAAAAATTAGGAAATTGTGTTTT---AATTCATGACTTT
LJFgene1       GCATTGGCCGGTGAAAGTAGTGCAATCTTTACTTTTTCTTTTTTTGAAATTTTTTTTTTC
                  **      * ***  * *  **  **      ** * ****   ****  *   **

LJFgene8       A--AGTTAAAGTCACGAATTAAAAGATGGCTTTAATAAAAGTAAA--TTTA---------
LJFgene1       ACTACCTTTGGTCCTTGTATTGAGCATGGTCCCACCAAAATCCAAACTTTATTATTGGAC
               *  *  *   ***      *  *  ****    *  ****   **  ****

LJFgene8       AAAAATAAATAAAGAAGTCA---TGTTTGAAAATATACCTT-TCGTTGAAATGTTGTGCT
LJFgene1       ATAGATGAATCATGATGTCACTTTGTTTTAATATTTTCATTCTTTTTCAGGAACTATAAT
               * * ** *** * ** ****   ***** ** ** * * ** *  ** *     * *  *

LJFgene8       TTTAAA------ACACAACTTGCT-------------CATTTATTGATTTTTTAAA----
LJFgene1       CCTGAAGGTAGGAAAAAACCTGTGAGCAGGGAAGAAGCAATGGAGGAACTTATAGACGTG
                 * **      * * *** **               ** *    **  ** ** *

LJFgene8       --AATAAATTTACCTATTTTGGGATTTTTTTTTTATAAAAAATGCACCACTCTAGTAAGT
LJFgene1       GTAATAAATGCAACTGAACTGAAATCTTGAGTTATCACTGGATTGAAATTTTCTTTTCCT
                 *******  * **    **  ** **   **   *    **  *    *    *   *

LJFgene8       AAACCTTTTTTTTATATATAAAAAATTGCAT-TCAGGTAAACAACATAATTAAGTGTTTA
LJFgene1       TCCCTCATTTTATCAACATTGATTATAATTTATAAAATTTATGAAAGGAGTGATAGTGAA
                  *   **** *  * **  *  **    * * *  *  *  * *  * * *  **  *

LJFgene8       CTGATTCATTGAAACACTTATGTATAAGTTGTTTATGTGATTGAAGAGAAAATAAAGTTA
LJFgene1       TATACACTTTGTAACACTATTTCTAATACACTTTCTATTATCGGTTAAAATTTATTGAAA
                  *  * *** ******  *    *     *** * * ** *   * **  **  *  *

LJFgene8       A-ATTATTTTCTTATAAATTGTAATATGTTTTCATG----AGCTATGGAA-AGTTTATTG
LJFgene1       ACAGTGTTGTTAAATGGCGGCCATGGCGGCGCCATGGCGGAGTTGCGTAACGGTTTTCTG
               * * * ** *   **       *    *    ****    ** *  * **  ****  **

LJFgene8       AAATAA----ACTGAAAATAGATTGTG-GATATTTCATAAATAC--ATATCTTAA-AATT
LJFgene1       AAAAAACGCCACCGAATAACGGTGGCGTGGCGGATTAAAGATGGCGGCGCCATGGCGGCC
               *** **    ** *** *  * * * * *     * * * **        * *

LJFgene8       ATTTTAAATATTCCAAACACTTATATAAATACTATAAGCACTTGTAATAGAAGAAA----
LJFgene1       GCCATAGCCATGGCGGCCA-TGGCGGATGTGGCGGGGAGGCGGAAAATGGCAGAATTTTT
                   **   **  *   ** *     *  *          *    *** * ****

LJFgene8       ------------GATAAGAAT---GTAAAATAAATTGATTTTTTTTCCATAAGTTGATTT
LJFgene1       TTTTTTGTCCGCGGTAGGAGTTGGGCTGACCCGATCCAACCCTACCCGAAAACTTAATGA
                           * ** ** *   *   *    **  *    *   * * ** ** **

LJFgene8       AAGTTAAAATCAACT-TATGTATCATAACACCTCAG--GTTTTGAAAAAGTTAAATGAGA
LJFgene1       AAACCATGACCCCCCCTACCTTTCAGAACGCTGCTGCAACCCTCGAAGCTTCAACATAGC
               **   *  * *  *  **  * *** *** *  * *      *  **   * **   **

LJFgene8       GAGTTTTT--AACAAAGTTAAGTGTATAAGTTAAATTTAAGAGAAAAACACAATTTATTC
LJFgene1       GCGACCTCGGAACCAGCCACGACCCCTGCAACCAGCCTCGGAACCAGCCGCG-CTTGACC
               * *   *   *** *           *      *   *  **   *  * *   **   *

LJFgene8       TACCTTTTCTTTTCCCCATTTGTAATTGTTTATGGCCAATTTGATCCAAACAACATCTTT
LJFgene1       CGCTGCACACAGCAGCCAGCAGCGACGACCCATGGCG---TGAACGGCGGCGACGAGCAC
                 *            ***   *  *      *****    *  *      * **

LJFgene8       AGCCTAAATAAGCTCTTC-CAATCACACTCTAAGTTTAAGTATGATTACTATGATGTTTA
LJFgene1       GGCGTGAACAACGGCGACGCGAACGGAGAAGACGACCCAGAACCGCGACCTCGACTTATA
                ** * ** **   *  * * * *  *    * *    ** *     **   **  * **

LJFgene8       GATTTCATTGTGTTG-----TTTTTTATGACCTTAATTTTCCGTCGAGTAACGCCATATT
LJFgene1       CGGGTGGGTGGGCCAAAGCCTTTTTTTTGTTTTTGCTGCTTGACACCCCTTTTTTATGTC
                   *   ** *        ****** **   **  *  *               ** *
```

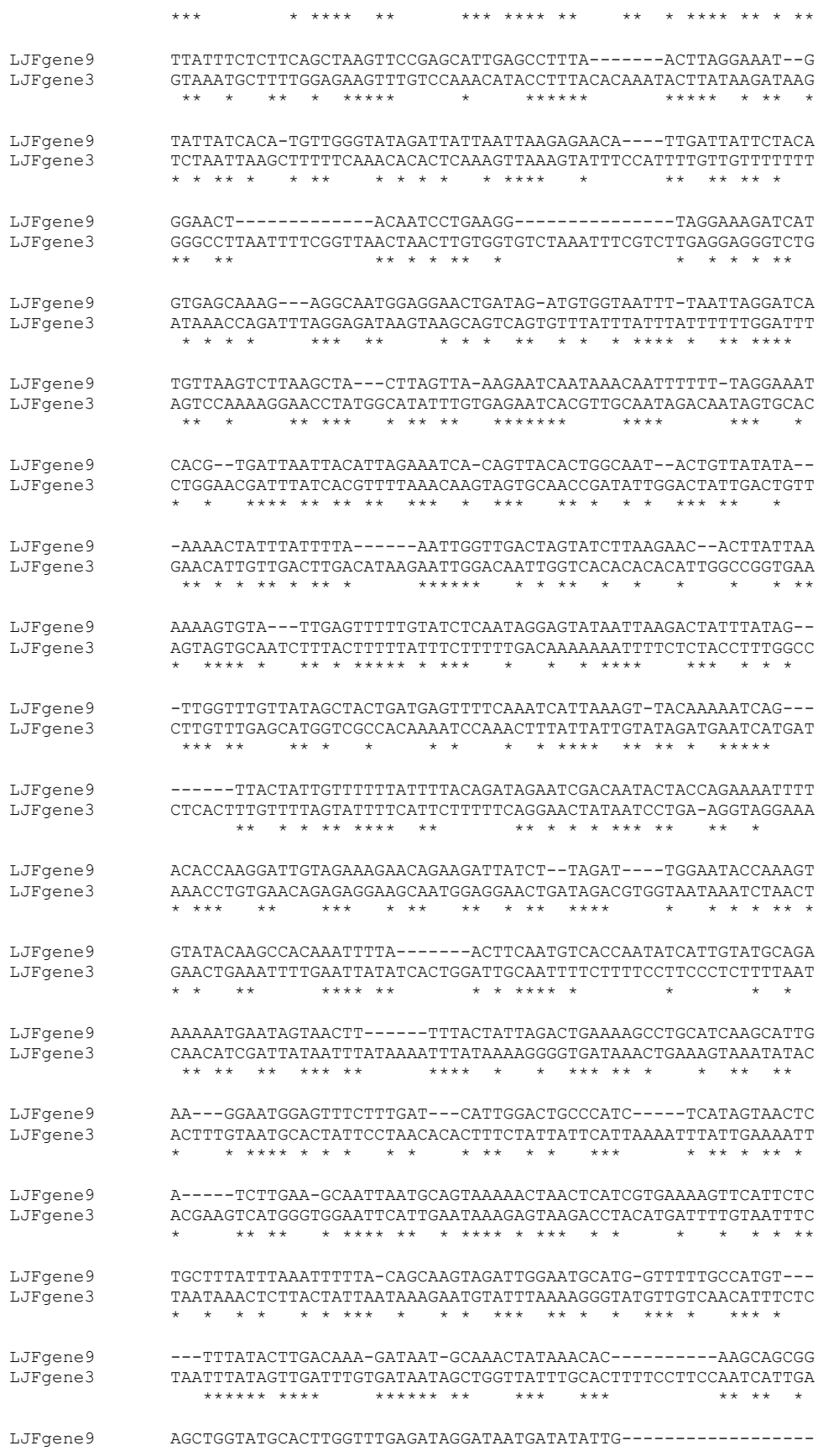
```
LJFgene8       TGAATTAAAGTTTGTAAATTT-TAGATGAAATTCA-TTTTGCAAACTGAATTTGTAAAAT
LJFgene1       TGCAGCTTTTTTTCCGTTTTTCTATTTGACACCCCCTTTTACTTTCGACAGTCCCATTTT
               ** *      ***     *** **  *** *  *  **** *   *   * *   *   *

LJFgene8       CAACTCTTCCTTTACTTCCAAAATCCAAACTTTATTATTGGACAAAGATGAATCGTGGAG
LJFgene1       TAATTTTTTTTTCTATTTGACACCACAATTTTTTTTTCTGTTCAGTCCTCTTTTAAATGG
                ** * **  **   **  * *   ***  *** **  **  **    *   *      *

LJFgene8       TTACTTTGTTTTAAATTTT--TAATTTTAATTATTGTTCAGGAACTACAATCCTGAAGGT
LJFgene1       CTGAACCATCATCCTCTTTAATGAGTTTATTTGGTGTTGGTACTTGATTATTGTATGAAC
                *      *  *    ***  * * **** **  ****        *  **  *

LJFgene8       AGGAAAAACCCTGTGAGCAGAGAAGAG-GCAATGGAGGAACTGATAGACGTGGTAATAAA
LJFgene1       TCATGAAACTTTTTTAGTTTATTTGAATGCAATCCTTTGTTTTTTTCA-ATTTCAATGAG
                    ****  * * **   *   **  *****        *  *  *  *   *** *

LJFgene8       TCTAGCTGAAATCATAAGTTATTTTCATGAATGCATTGCAATTTCCCTTTCCTAGCCTGT
LJFgene1       TTTA--TATATATGTTTTTTTTTTTTTTGGTCCGCCATGACTTCCGCCATTTT-CCGCT
               * **  *  *    *   ** ****  *   * *      * ****      *  **  *

LJFgene8       GTATCAACAAATGTTATTATTTATAATAAGTTTAATTTTCATGCACTGACCGTATATAAT
LJFgene1       ACGCCATCCGCCATATTTTTATGGCGGATTTTTTACTTTCCGCCATGAACCGCCATCCGC
                   ** *     *  ** * *     *  *** * ****   **   ****

LJFgene8       AATT--TTATATTGATATCTAATCATAAA-TCATCATTTAAATTATTTTAAGATAATTAA
LJFgene1       CATTAACAACATTGATTGAAAATTATGAAGTCATGAGAGGAGCTCATTGAA--TAAAGAG
                ***    * ******    *** ** ** **** *    *  *  ** **  ***  *

LJFgene8       TTTAAAAGTTAATAAATTTACCGTATATAATGAATTATAATTGAATAATTGTATAAAAAA
LJFgene1       TGAGAACTTACATGATTTTGTAATTTCTAATAATTTTTTACT--ATTAATAAAGAGTGTA
               *   **  *  ** * ***    * * **** * ** * * *  ** * *  * *    *

LJFgene8       TTTATAATGTCAC-TGTATAATTCTTTTTCTCATTTATAATTTGGTTGTAAGTTGTAGCT
LJFgene1       TTTAAATGGGTATGTTTTTGAACATTTTTCTAATTTCTAATCGATTTGTAA---------
               **** *  *  *  * * * *   ******* **** ****    ******

LJFgene8       TGTAATAGCAGGTTATTTACCTTTTTCCAATCATTGAAGTAAAGTTAAAATCCAGCTCTG
LJFgene1       --TAATAGCTGGTTATTTGCACTTTCCCAATCATTGAAGTAAAGTTAAA--CCAGTTCCG
                 ******* ******** *  *** **********************  **** ** *

LJFgene8       GATAATTTTATAGATAGAATCAACAACACCAGACAAATTTACACCACGAATAGTTGAAAG
LJFgene1       GATAATTTTACAGATAGAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAG
               ********** ***************************** ******* ***********

LJFgene8       GAAGGAAGACTATATTCATGTGGAGTACCAAAGCTCAATCTTGGGGGTATGTGTAACTTA
LJFgene1       GAAAGAAGACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGGTAAGTGTAACTTA
               *** ************* ******************************** **********

LJFgene8       CATCAAAAGGAAACTCATCGTGGAGAAAAATAATAATTTTGTACATTTTAGATGATAATC
LJFgene1       CATCTAA-GGAAACTCATCATGAAGAAAAATTATCCTTTTATACATTTTAGATGATA-TC
               **** ** *********** ** ******** **  **** *************** **

LJFgene8       AAGAACCATCTCTAATCCCTTCTCCCCTCCTTTTTATTTTTTCTGCCATGTGCTAGCAGT
LJFgene1       AAGATTCAAGAACCATCTTTAATTTCCTTCTCCCT--TTTTTCTGTCATGTGCTAACAGT
               ****  **      ***  *  *  *** **   *  ******** ********* ****

LJFgene8       TTGTGCATGATGTTGAGTTCTGGTTTCCACTGGGTAAGGGTTCTACTGTGGAGTATCGAT
LJFgene1       TTGTGGATGATGTTGAGTTCTGGTTTCCTCCGGGTAAGGGTTCTACTGTGGAGTATCGTT
               ***** ********************* * ************************** *

LJFgene8       CTGCATCTCGGTTGGGGAACTTTGATTTTGATGTGAATAAGAAAAGAATAAAGGTATGTT
LJFgene1       CTGCATCTCGGTTGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATGAT
               **************** ******************* *  ***************** *

LJFgene8       TGTATCATTCCTTTGTGCTGTCTCGGTAGTTAACATGAAGAAAT----GATTAAAAGATA
LJFgene1       TCCATAATTCATATGTGCTTTCTCTATAGTTAG-ATAAAGAAATTCTTGGTTCCAGGGTA
               *  ** **** * ****** ****  ******  ** *******    * **  * * **

LJFgene8       TTTTGTCCTTTAGGTTTTTTGGTTATATTTAGTTTGATTTTTTATTTTTTAAAAGTTAAA
LJFgene1       AAACTCCCCTT--TCCTTCATGTCATGTGAAGCATTTTTTACTCAAGTAGATCCACTAAA
```

```
                    ** **     **   ** ** *  **  *  ***  *    *  *     ****

LJFgene8        TTTAGTCCTTTATGTTTTTAAAACGAATCAAAATGATC--ACTATTTCGAGATGTAAAAC
LJFgene1        TTTGAGTCTCAAATGTTTTAACTTTATTCTAAATGTTTTAACTTTATTTGAGTCTCAAAT
                ***    **  *   ******    * ** ***** *   *** * *     * * ***

LJFgene8        TTTTTTATT------AGATTTACTTACATGCAAATTATGCAACTTGACTAATGTTTGAAA
LJFgene1        GTTTTAACTGAAGGTAAATTTGGTTAACTATGATCAGAAATACATTAACAAGAGTTGAAG
                 **** * *      * ****  ***  *   *        ** * *  **   *****

LJFgene8        AAAAAAATCATCAATATGATAAAAATAATGAGTGTCACAAGCAATTAAAAA---------
LJFgene1        AATATT-TTTTATGTACAATAAAGAGAGTATTTGCTCGAGAGAATGTAAATCCTTTTCTA
                ** *   *  *   **  ***** * * *   **    *   ***  ***

LJFgene8        ------------------------------------------------------------
LJFgene1        AATATTTTTGTGATGAAAAATAATGGTTGCTGGCAGGCACTGAGACAAGAGTTGGAGAAG


LJFgene8        ------------------------------------------------------------
LJFgene1        AAAGGATGGGCATCTCAAGACACCATATGATGAAAAAACTTAGGCAGAATTCACATCAGC


LJFgene8        ------------------------------------------------------------
LJFgene1        ATCTAAGAAAATATTGTTTCATATACATTGTAACCTTGTATACTTTTGTATTAGATACAA


LJFgene8        ------------------------------------------------------------
LJFgene1        AATCTCACAAGATCATTGAAAGCAAACTCTTCATGATTATTGGAATTGTAGAAATGATTG


LJFgene8        ------------------------------------------------------------
LJFgene1        AGAACAGTACTTCAAACTCTCGGGGGAAGGAATGAAATGAAGATGTTACCCATATCTTTT


LJFgene8        ------------------------------------------------------------
LJFgene1        TGAACTTCATTAATTGGTCCACTTATTTACTCTTTCGCTGAGTTCAATCTAACAATGTAG


LJFgene8        ------------------------------------------------------------
LJFgene1        CATTCTTGTTTCAAGAATTTTATGTTGTGTTGTTTTTCAAATGCATTTGCCATGATTCAC


LJFgene8        ------------------------------------------------------------
LJFgene1        CTTCGGTGCATATAGCATTTGGAATCTGATTATAATAGGAGCATTGCGGAGACCAAAAAC


LJFgene8        ----------
LJFgene1        AGGGTGCATT
```

LJFgene9 vs. LJFgene3

```
LJFgene9        ------------------------------------------------------------
LJFgene3        CCATAAAAAAGAAAAAAAAAAAGTCCCACCGCCCACCTTCTTTATCACATGATTCACATC


LJFgene9        ------------------------------------------------------------
LJFgene3        TCATTCCTTATATTTGGTTCACATTCTTAAATTATAAATATTTCGGTCTGTGAAGATATA


LJFgene9        --------------AGCATTGGTTCGCACCTAAGGCACC--TTCCCAATTCAGCTTCTAA
LJFgene3        TGTCCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAACAACAATGGCTT
                              ***   *** ****       *   ******      *

LJFgene9        CGATGACACTTTGTAGCACGTTTTCCAACTTCAA---CATTCACATATTAAAAAACAACA
LJFgene3        CAATGGCATCTTCAAGCTCCTTCTGCAACCTCAAGTTCATCACCAAACCCAACAATGGTA
                * *** **  **  *** * ** * **** ****   ***   ** *   ** **    *

LJFgene9        AGGGTTCCTTTTCTCGTCGATTTCAACTCTCTCAGAAGCTGGATGACGATAATTTCATTG
LJFgene3        GAAGAAGCTCTCTTCCCCGTATTGTATTCTGTCAGAAGCACCACGATAGCACACCCACCG
                   *   ** *  **  **  **  * *** ********   * **    *    **  *
```

```
LJFgene9       ATAAAATCAAACGAAGGTTCTCACTGATTCTCCCTTTAATTTGCCA----CCTCACATGA
LJFgene3       ACCAAATCAACCGAAGGTTCTTATTTCTTCACACTCGCACTTTCTAATTCCTTTCTATGG
               *  ******* ********** * *  *** * **   * ** * *    * *   ***

LJFgene9       ATTGTATAATATATATTTATATTTATGCT------------TGACCTTGAATTGTTCC-T
LJFgene3       ATTAT-TCATATCTATTCATACCCATCTTCTGAAATCTCTTTATATTTCAATTATTTTGT
               *** * * **** **** ***   **  *            *    ** **** **   *

LJFgene9       ATCTTAAAGAGAGCTCATACTGGAAAGTGGAGAATTAGCAACCATTGGTGCCATCTTCAA
LJFgene3       CTATTGAAGAGAACTCATATTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTGAA
                * ** ****** ****** **  ***  * *** **** ***************** **

LJFgene9       CTTTAGGTACACTGCTTTATTGTTTTCACAATGAGAATAGGTGACCAAAATTTAATTAAT
LJFgene3       CTTCGGGTACCCCTCCTCTGCTTGTTTTTGG---------------AAAATTT--TTGTT
               ***  ***** *  * *     * **                    *******  **  *

LJFgene9       TCTTATAATATTTGAAAATATTTCGAGAAATTTTATGTAAAATTAACTCTTTTTCTTTAA
LJFgene3       TTTCATTTTATTTTGAA-TGT------AAATTGAATTCAAGATTTGATTTTGTTGGT--G
               * * **  *****  ** * *      *****  **  ** ***   * ** **  *

LJFgene9       GATTTATGTGATTACTTCATAGCAAATTCTTGTCAAGTTTTAAGAACTTTTGAGTTTTGA
LJFgene3       GGTTTGAAGACCCTTTTGGTTTTTAATTTCGGTTTTGTTTTG-----TATTGGACATGGG
               * ***          **  *    ****   **   *****      * ***    * *

LJFgene9       TGTGTGTTTTCTATTGGAAATATGTGGGGATATATTATTTTGGATTTTATGCTTCTTTGC
LJFgene3       TG-GTGGTTAAAAAAGAGAAAAT-TGAGTTTGTGT--CTTGTGTTTTGATGGT-----GC
               ** *** **   *  *  ** ** ** *  * * *   **  * *** *** *     **

LJFgene9       AGAGGCAAAAAGCCAGATTATCTTGGAGTGCAGAAAAATCAACCGGCATTAGCACTATGT
LJFgene3       AGTGGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCAGCATTAGCTCTGTGC
               ** ** ***** ** ********************** * *** ******** ** **

LJFgene9       CCGGCAACTAAGAACTGCATATCGACATCTGAAAATGTCACTAACCTCACACATTACACT
LJFgene3       CCGGCAACGAAGAATTGCGTGTCAACCTCTGAGAATATCAGTGATCGCACACATTATGCT
               ******** ***** *** * ** ** ***** *** *** * * * *********  **

LJFgene9       CCTCCTTGGTGAAATTTCCTTCTTTATTTTTTTATTTATTAAAGTTTTAACTTTGGTTTA
LJFgene3       CCTCCATGGTAAAAGTT--TCCTTCTTTTTCTTATTTT---AATTTTCACCTTGGATTTA
               ***** **** *** **  * ***  **** ******    ** *** * *** * ****

LJFgene9       TGATATTATCTGAATC---TGAATTGGCTGCAAAGCCTG------CAATTTGTT--ATTG
LJFgene3       TGGGATTATATGTATTAAATGCATTTTTTTAATTGTGTGTTTGGACAAACTACTTAGTTA
               **  ***** ** **    ** ***   *  *  *  **      ***  *  *   **

LJFgene9       AATATATATTGTTGTTGAGTTTAATTTTACAC-TGAATAACTTTTAAAATAATTATTATA
LJFgene3       AGTGCTCATC-TCATCATGTAAGTGCTTATGCATAAGTTGTTTCTATAATAAAAAA-ATA
               * *    **  *  *   **      ***  * * * *   ** ** *****  *  ***

LJFgene9       AAAATCAATAAAATTA--TTATCTATATAAAAAATATGATTAGATGATAAAGTAAAACTT
LJFgene3       AAAATACATACGTATGAGTTGTTGTTGTAAGCTTTTTTCTTAAGTTATTCTGGAAATCTT
               *****  ***    *   ** *   * ***    * *  ***  * **   * *** ***

LJFgene9       TTGTATGGTATTAATACATGAATTATTTTTTGTCCAGAATTTTTCAGATTG--TCATAAA
LJFgene3       ATTGAAATAATCTGAAAACAACTTTTTTTTTTACATGATCTGCTGAAAACAACTTATAGA
                *  *    **    * *  * ** ******  *  **  *  * * *     * *** *

LJFgene9       T---------TTGCATTTTGATAAAGTGTATAATTC-ACTCAATTTCAAAATAACTGACT
LJFgene3       CATATGGTAATCACATATCATTAAGTTAAGTTATTCCAAACACTTACATAAACACTTATA
                         *  *** *   ***  *   * **** *  ** ** ** **  *** *

LJFgene9       TTTTAACACCAATTTGG--------CAAAATGTTGCTTATGGAACAATTTAATTAATT-A
LJFgene3       AGAGAAAAACAATAAGAAATAAAAACAAAATAAAATTTTCAATAAGTTAAAATTAGTTTA
                   ** * ****  *         ******     **     *   *  ***** ** *

LJFgene9       TCACATTTTAAAACCATTG--AAAGTTCAATTGGTGTTGCCGTGTTGGGCGCATTGAGCC
LJFgene3       TAAAAGTTTTTTTTTATTATAGAAGCTCTCTTGTATTAGCCTCTCCTAAAGTTTTTTTTT
               * * * ***      ***    *** **  ***   * ***         *  **

LJFgene9       CTT--------AACTTAGGAAATGTATTATCTCATGTTGAATTTGCTTTTTTTTTTTTTT
LJFgene3       CTTCATAATTTAGCTTAAAAAGAAACCTATTTCATTTTTTCATT-TTATTTTCTTCTCTT
```


LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9
LJFgene3
LJFgene9

```
LJFgene3       AGTAAAGTTAAACCAGTTCCGGACAATTTTGCAGATAGAATCAACAACACCAGACAAATT
               **      *  **  * *  *    *   *   **** * *

LJFgene9       ------------------------------------------------------------
LJFgene3       TTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAGTACCAAAGCTCAAT

LJFgene9       ------------------------------------------------------------
LJFgene3       TTTGGGGGTAAGTGTAACTTACATCTAAGGAAACTCATCACGAAGAAAAATGATAATTTT

LJFgene9       ------------------------------------------------------------
LJFgene3       ATACATTTGAGATGATATCAAGATTCAAGAACCATCTCTAATTTCCTTCCCTCCTTTTTC

LJFgene9       ------------------------------------------------------------
LJFgene3       TGTCATGTGCTAACAGTTTGTAGATGATGTTGAGTTCTGGTTCCCACCGGGTAAGGGTTC

LJFgene9       ------------------------------------------------------------
LJFgene3       TACTGTGGAGTACCGATCTGCATCTCGGTTAGGAAACTTTGATTTTGATGTGAACAGAAA

LJFgene9       ------------------------------------------------------------
LJFgene3       AAGAATAAAGGTGTGATTTCATAATTCATGTGTTTTCTCTATAGTTAGATAAAGAAATTC

LJFgene9       ------------------------------------------------------------
LJFgene3       TTGGTTCCATGGTAAAACTCCTCTTTCCTTCATGTCATGTCAAACATTTTATACTCAAGT

LJFgene9       ------------------------------------------------------------
LJFgene3       AGATGATTCACTAAATTTGAGTCTCAAATGTTTTAACTTTATTCTAAATTAGTCACTTAT

LJFgene9       ------------------------------------------------------------
LJFgene3       TTTAACTGAAGGTAAATTTGGTTAACTATGATCAGAAATACATTGACATTTTTTAATTGG

LJFgene9       ------------------------------------------------------------
LJFgene3       TAGAGATAAAGAATATTTTTTATGTACAATAAAGAGAGTATTTACTCCAGAGGATGTAAA

LJFgene9       ------------------------------------------------------------
LJFgene3       TCCCTTGCTAAATATTTTTGTGATGAAAAATCTTGGTTGTTGACAGGCACTGCGACAAGA

LJFgene9       ------------------------------------------------------------
LJFgene3       GTTGGAGAAGAAAGGATGGGCATCTCAAGACACCATATGATGAATAAACTCAGGCAGAAT

LJFgene9       ------------------------------------------------------------
LJFgene3       TAACATCAGCATCTAAGCAAATATTATTTCATATACTTTGTGACCTTGTATACATTTGTA

LJFgene9       ------------------------------------------------------------
LJFgene3       TTAGATACAAATCTCACAGGATCATTGAAAGCAAACTTTTCTTTGATTATTGGAATTGTA

LJFgene9       ------------------------------------------------------------
LJFgene3       GAGAAATCATTGAGAACAGTACTTCAAACTCTCGGGGAAGGAATGAAATGAAGACCTTGC

LJFgene9       ------------------------------------------------------------
LJFgene3       CCCATATCCTTCTTCAAGTTCATTAATTGGTCCGCTTATTTACTCTTTCACCAAGTTCAA

LJFgene9       ------------------------------------------------------------
LJFgene3       TCTAACAATGTATCGTTCTTGTTTCAAGAATATTAATTTGTGTTTTGTTTCTAATGTGTT

LJFgene9       ------------------------------------------------------------
```

```
LJFgene3       CTTCACGATTCACTTTTGTTGTAAGTTTGACTTAGCTTCTTCATGTATTAAAACTTAACC


LJFgene9       ------------------------------------------------------------
LJFgene3       TTGTGATCAATTTGGATACTTTAGGCCTTTCTCTTATCAAAATGTACCTATGTCATGATC


LJFgene9       ---------------------------------------------
LJFgene3       TCGAAATTGTTCAATCTCATTGATCCTAACTAGAAAATTCGTCCAA
```

LJFgene9 vs. LJFgene14

```
LJFgene9        ------------------------------------------------------------
LJFgene14       AAGTCCCACCTTCTTTTATTCATCACATGATTCACATCTCATTTCTTATTTTCGGGTCAC


LJFgene9        ------------------------------------------------------------AGCAT
LJFgene14       TTTGTTAAATTATAAATAATTTCGTTCTGTGAAGGTACACACGTTCATAAGTTCCTTAAT
                                                                            **

LJFgene9        TGGTTCGCACCTAAGGCACC--TTCCCAATTCAGCTTCTAACGATGACACTTTGTAGCAC
LJFgene14       TTTCTCGAACCTTCATTTTCAGCTCCCAACAATAATGGCTTCAATGGCATCTTCAAGCTC
                *   *** ****       *   ******      *     * *** **  **  *** *

LJFgene9        GTTTTCCAACTTCAA---CATTCACATATTAAAAAACAACAAGGGTTCCTTTTCTCGTCG
LJFgene14       CTTCTGCAACCTCAAGTTTATCACCAAACCCAACAATGGTAGAAGAAGCTCTCTTCGCCG
                 ** * **** ****    **   ** *   ** **    *   *   ** *  *** **

LJFgene9        ATTTCAACTCTCTCAGAAGCTGGATGACGATAATTTCATTGATAAAATCAAACGAAGGTT
LJFgene14       TATTGTATTTTGTCAGAAGCATCACGATGACACACCCACCGACCAAATCAACCGAAGGTT
                  **  * * * ********   * ** ** *    **  **  ******* ********

LJFgene9        CTCACTGATTCTCCCTTTAATTTGCCACCTCACATGAATTG------TATAATATATATT
LJFgene14       CTTACTTCTTCACACTCACACTTTCTATTTCCTTTCTATTGATTATTCGTAACCATCTTC
                ** ***  *** * **   * ** * *  **   *  ****        ***      *

LJFgene9        TATATTTATGCTTGACCTTGAATTGTTCCTATCTTAAAGAGAGCTCATACTGGAAAGTGG
LJFgene14       TGAAATCTCGTTACATTTCAATTCTTTTGTGTATTGAAGAGAACTCATATTGAGAAGCAG
                *  * *   * *  *  *  * *  **  * * ** ****** ****** **  ***  *

LJFgene9        AGAATTAGCAACCATTGGTGCCATCTTCAACTTTAGGTACACTGCTTTATTGTTTTCACA
LJFgene14       CGAAATAGCGACCATTGGTGCCATCTTCAACTTCGGGTACCC--CTCCTCTGTTTTTGC-
                 *** **** **********************  ***** *  **    ******  *

LJFgene9        ATGAGAATAGGTGACCAAAATTTAATTAATTCTTATAATATTTGAAAATATTTCGAGAAA
LJFgene14       -TCTGTTTTTTTTTCTGGAAATT--TTAGTTTTT----CATTT-----TATTTTGA----
                 *  *  *   *  *   ** **  *** ** **     ****     ***** **

LJFgene9        TTTTATGTAAAATTAACTCTTTTTCTTTAAGATTTATGTGATTACTTCATAGCAAATTCT
LJFgene14       ----ATGTAAATTAAATTC---------GAGATTT----GATT--TTGTTAGTGGGTG-T
                    ******* * ** **          ******    ****  **  ***    *  *

LJFgene9        TGTCAAGTTTTAAGAACTTTTGAGTTTTGATGTGTGTTTTCTATTGGAAATATGTGGGGA
LJFgene14       TGAGACCCTTTTGGA---TTTTAGTTTGGGT-TGTGTTTTGTATTGGAAATGGGTGG---
                **  *   ***  **   *** ***** * * ******** **********  ****

LJFgene9        TATATTATTTTGGATTTTATGCTTCTTTG---CAGAGGCAAAAAGCCAGATTATCTTGGA
LJFgene14       --------TTTGGGTTTTGTGTTTTGGTGGTGCAGTGGGAAAAAACCTGATTATCTTGGA
                        ***** **** ** **   **   *** ** ***** ** ************

LJFgene9        GTGCAGAAAAATCAACCGGCATTAGCACTATGTCCGGCAACTAAGAACTGCATATCGACA
LJFgene14       GTGCAGAAAAACCCACCAGCATTAGCTCTGTGTCCGCCAACTAAGAACTGCGTGTCAACC
                *********** * *** ******** ** ****** ************** * ** **

LJFgene9        TCTGAAAATGTCACTAACCTCACACATTACACTCCTCCTTGGTGAAATTTCCTTCTTTAT
LJFgene14       TCTGAGAATATCAGCGATCGCACACATTATGCTCCTCCATGGTAAAAGTTCCCTTCT--T
                ***** *** ***   * * *********  ******* **** *** **** *  *  *

LJFgene9        TTTTTTATTTATTAAAGTTTTAACTTTGGTTTATGATATTATCTGAATCTGAATTGGCTG
LJFgene14       TTTCTTATTTT---AATTTTCACCTTGGATTTATGGGATTATATGAAT-TGAATGCAATT
                *** ******    ** *** * *** * ******  ***** ***** *****    *
```

```
LJFgene9         CAAAGCCTGCAATTTGTTATTGAATATATATTGTTGTTGA--GTTTAATT--TTACACTG
LJFgene14        TTTTAATTGTGTTTGGATAAACAACTTAGTTAACTGCCCATCATGTAAGTGCTTATGTAT
                        **   ** * **   **  **  *   **   *   * *** *  ***

LJFgene9         AATAACTTT-TAAAATAATTAT-TATAAAAATCAATAAAATTATT----ATCTATATAAA
LJFgene14        AAGGTTTTTCTATAGTAAAAAAATGTACAAGTTGCAAGCTTTTTTCTTAATTTATTCTTG
                 **    *** ** * ***  *  * ** ** *    *   ** **    ** ***

LJFgene9         AAATATGATT-AGATGAT--AAAGTAAAACTTTTGTATGGTAT---TAATACATGAATTA
LJFgene14        AAATCTTATTGAAATAATCTGAGAACAACTTTTTCTTTTTTACCTGTGATCTGAAAACAA
                 **** * *** * ** **   *    **  **** * *  **    * **     **  *

LJFgene9         TTTTTTGTC---CAGAATTTTTCAGATTGTCATAAATTTGCATTTTGATA-----AAGTG
LJFgene14        CTCATAGACATATCATAAACTTGGGATATAGATAAATTTTTTCATAAACACTTACAAGAG
                  *  * * *       *   **  ***    ********     *  * *     *** *

LJFgene9         TATAATTCACTCAATTTCAAAATAACTGACTTTTTAACACCAATTTGGCAAAATGTTGCT
LJFgene14        AAAAAAAAATACAAAAGAAAAATAAAATAAATTTTTCTATAAGCTAAAATTAGTGT-ATG
                  * **   *  ***    *******   *  ****   *  *  *      * ***

LJFgene9         TATGGAACAATTTAATTAATTATCACATTTTAAAACCATTGAAAGTTCAATTGGTGTTGC
LJFgene14        TGTAAGCTAATTTG--TAGTTCTTTCATATTAGCTTCTCCAAAAGTTTTTTTTTTTTTTA
                 * *     *****   ** ** *  *** ***    *    ******   **  * **

LJFgene9         CGTGTTGGGCGCATTGAGCCCTTAACTTAGG---AAATGTATTA-----TCTCATGTTGA
LJFgene14        ACTTATGCATAAGTTAAA--TTTAGCTTAAAGATAAATTTATTTCATTTTCTTCTCTTGT
                   *  **      ** *    *** ****     **** ****      ***  * ***

LJFgene9         ATTTGCTTTTTTT----TTTTTTTTTATTTCTCTTCAGCTAAGTTCCGAGCA--------
LJFgene14        AAATGCTTTTGGAGAAATGTATCCAAACAGACCTTTACACAAGTACTAATAAGATAAGTC
                 *  *******       * * *    *     *** *   **** *  *  *

LJFgene9         ---TTGAGCCTTT----AACTTAGGAAAT----GTATTATCAC-ATGTTGGGTAT--AGA
LJFgene14        TAATTAAGCCTTTTCAAAACGCTCAAAGTTCAAGTATTTCCATTATGTTGATTCTTCGGG
                    ** *******    ***     ** *    *****  **  ******  * *   *

LJFgene9         TTATTAATTAAGAGAACATTGATTATTCTACAGGAACTACAATCCT--GAAGGTAGGAAA
LJFgene14        CCTTAATTTTCGGTTAAGTAACTTGTGGTGTCAAAATTGCGTGTTGAGGAGGGTCTGGTA
                    * * **  *   *  *   ** *  *     ** * *        ** ***  *  *

LJFgene9         GATCATGTG-AGCAAAGAGGCAATGGAGG---AACTGATAGATGTGGTAATTTTAATTAG
LJFgene14        AACCAGATTTAGGAGATAAGTAATCAGGGTTTATTTATTTATTGGATTTAGTCCAAACGG
                  * **  *  ** * * * * ***   **   *  *  *   **   * * *  **   *

LJFgene9         GATC-ATGTTAAGTCTTAAGCTACTTA-GTTAAAGAATCAATAAACAATTTTTTTAGGAA
LJFgene14        AACCTATGGCCATATTTGTGAGAATCACGTTGCAATAATAGACAATGGTGCACTTGGAAC
                  * * ***   *   **  *  * * * ***  *  *  *   **   *    ** * *

LJFgene9         A-----TCACGTGATTAATTACATTAGAA-ATCACAGTTACACTGGCAATACTGTTATAT
LJFgene14        AATTTATCACGTTTTAAACCACGTGAAAGTAGTGCAGCCGAAATTGGACTATTGACTGTT
                 *     ******  * **  ** * * *  *   ***    * * * * ** **     *

LJFgene9         AAAAACTATTTATTT----TAAATTGGTTGACTAGTATCTTAAGAACACTTATTAAAAAA
LJFgene14        GAACAATGTTGACTTGACATGAATTGGACTATTGGTCACACACATTGGCCGGTGAAAGCA
                  ** * * ** * **    * ******   * * **  *  *      *   * ***  *

LJFgene9         GTGTA---TTGAGTTTTTGTATCTCAATAGGAGTATAATTAAGACTATTTATAGTTGGT-
LJFgene14        GTGCAATCTTAACTTTTTCTTTTTTTTTTGACAACTTTTTTTTTTTCTCTATCTTTGGTC
                 *** *   ** * ***** * * *   * *     *  **     * * ***  *****

LJFgene9         -TTGTTATAGCTACTG----ATGAGTTTTCAAA-----TCATTAAAGTTACAAAAATCA-
LJFgene14        CTTGTATTGAGCATGGTCCCACCAAAATCCAAACTTTATTATTGGACATAGATGAATCAT
                  ****  *    *  *    *  *   * ****     * ***  *  ** *  *****

LJFgene9         GTTACTATTGTTTTTTATTTTACAGAT--------AGAATCGACAATACTACCAGAAAAT
LJFgene14        GATGTCACTTTGTTTTAATTTTTAGTTTCTTTTTCAGGAACTATAATCCTGA-AGGAAGG
                 * *   * * * ***** ***  ** *        ** * * * *** **   ** **

LJFgene9         TTTACACCAAGGATTGTAGAAAGAACAGAAGATTATCT--TAGAT----TGGAATACCAA
LJFgene14        AAAAAACCTGTGAGCAGAGAGGAAGCAATGGAGGAACTGATAGACGTGGTAATAAATCTA
```

```
                    * ***   **    ***   * **   **  * **  ****     *   * * * *

LJFgene9        AGTGTATACAAGCCACAAATTTTAACTTCA-ATGTCACCAATATCATTGTATGCAGAAAA
LJFgene14       GCTGAACTGAAATCTTGAGTTATAACACTAGATTGCAATTTTCTTTTCCTTCCCTCATTT
                  ** *   **  *   * ** ****   * **  **    * *  *  *   *  *

LJFgene9        AATGAATAGTAACTTTTTACTATTAGACTGA-AAAGCCTGCATCAAGCATTGAAGGAATG
LJFgene14       TATCAACATCGATTATAATTTATAAAATTTATAAAAGGAGTGATAGTAAATATACACTTT
                 ** ** *   * * *    *** * * * * ***    *    *   * *  *    *

LJFgene9        GAGTTTCTTTGATCATTGGACTGCCCATCTCATAGTAACTCA---TCTTGAAGCAATTAA
LJFgene14       GTAACACACTATTCTTAACCCCACCCTTTTTATTATTGGTTAAAATTTATCAGAAATTAG
                *     *  *  ** *    *  *** * * **  *   * *   * *   ** *****

LJFgene9        TGCAGTAAAAACT-AACTCATCG--TGAAAAGT----------TCATTCTCTGCTTTAT-
LJFgene14       AAAGTCATGAGTGGAACTCATTGAATAAAGAGTGAGACCTATGTGATTTTATAATTTCTA
                      *  *    ******* *  * ** ***          * *** * *  *** *

LJFgene9        -TTAAATTTTTACAGCAAGTAGATT----GGAATGCATGGT---TTTTGCCATGTTTTAT
LJFgene14       ATAAACTTATTATTAATAGTGTATTTAAAAGGATACATTGTGAACATTTCTCTAATTTAT
                 * ** ** ***     ***  ***     * ** *** **     ** *  *  *****

LJFgene9        ACTTGACAAA-GATAAT-GCAAACTATAAACAC------AAGCAGCGGAGCTGGTATGCA
LJFgene14       AATTGATTTGTAATAATAGCTGGTTATTTGCACTTTTCCAATCATTGAAGTAAAGTTAAA
                * ****      ***** **    ***   ***      ** **  * **      *  *

LJFgene9        CTTGGTTTGAGATAGGATAAT-GATATATTG-----------------------------
LJFgene14       ACTAGTTCCGGATAATTTTACAGATCGAATCAACAACACCAGACAAATTTTCACCACGGA
                  * ***   ****   * *  ***  * *

LJFgene9        ------------------------------------------------------------
LJFgene14       TAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGGTAA

LJFgene9        ------------------------------------------------------------
LJFgene14       GTGTAACTTACATCTAAGGAAACTCATCTTGAAGAAAAATGATCATTTTATACATTTGAG

LJFgene9        ------------------------------------------------------------
LJFgene14       ATGATATCAAGAACCATCTCTAATTTCCTTCTCCCTTTTTTCTGTCATGTGCTAACAGTT

LJFgene9        ------------------------------------------------------------
LJFgene14       TGTGGATGATGTTGAGTTCTGGTTTCCACCCGGTAAGGGTTCTACTGTGGAGTATCGATC

LJFgene9        ------------------------------------------------------------
LJFgene14       TGCATCTCGGTTGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATGATT

LJFgene9        ------------------------------------------------------------
LJFgene14       TCATAATTAATATGTGCTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCAGGGTAAA

LJFgene9        ------------------------------------------------------------
LJFgene14       ACTCCCCTTCCTTCATGTCATGTCAAACATTTTATACTTAAGTAGATTCACTAAATTTGA

LJFgene9        ------------------------------------------------------------
LJFgene14       GTCTCAAATGTTTTAACTTTATTCTAAATTAGTCACTTATTTTAACGGAAGGTAAATTTG

LJFgene9        ------------------------------------------------------------
LJFgene14       GTTGACTATGATGAGAAATACGTTGATATTTTTTAATTGGTAGAGATAAAGAATATTTTT

LJFgene9        ------------------------------------------------------------
LJFgene14       TATGTACAATAAAGAGAGTATTTACTCCAGAGGATGCAAATCCCTTACTAAATATTTTTG

LJFgene9        ------------------------------------------------------------
```

```
LJFgene14       TGATGAAAAATCTTGGTTGCTGACAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGA

LJFgene9        ------------------------------------------------------------
LJFgene14       CATCTCAAGATACCATATGATTAATAAACTCAGGCTGAATTAGCATCAGCATCTAAGCAA

LJFgene9        ------------------------------------------------------------
LJFgene14       ATATTATTTCATATACTTTGGACCTTGTATACTTTTGTATTAGATACAAATCGCACAGGA

LJFgene9        ------------------------------------------------------------
LJFgene14       TCATTGCAAGCAAACTTTTCTTAGATTTTTGGAATTGTAGAGAAATCATTGAGAACAGTA

LJFgene9        ------------------------------------------------------------
LJFgene14       CTTCAAACTCTCTCGGGGAAGGAATGAAATGAAGACCTTGCGCCATATCTTCTTCAAGTT

LJFgene9        ------------------------------------------------------------
LJFgene14       CATTAATTGGTCCACTTATTTACACCTTCACCGAGTTCAATCTAATAATGTATGAATCTT

LJFgene9        ------------------------------------------------------------
LJFgene14       GTTTCAAGAAATTTAATTTGTGTTTTGTTTCAAACGTGTTCTTCATGATTCACTTTTGTT

LJFgene9        ------------------------------------------------------------
LJFgene14       GTAAGTTTGACTTTGCTTCTTCATGTATTAAAGCTTATCCTTGCGATCAATTTGGATGAT

LJFgene9        ------------------------------------------------------------
LJFgene14       GCTTTAGGCCTTTCTGTCATCAAAATGTACGTACCTATGTCATGTTTTCAGAGGCTTTTT

LJFgene9        ---------------------------
LJFgene14       GATATCCTTGATGTCTTTACAAGAAA
```

LJFgene9 vs. LJFgene1

```
LJFgene9        ------------------------------------------------------------
LJFgene1        TTCCATATTTTCGGGTCACATTCTCAAATTATTATAACTAATTTCGTCATGTGAAGATAC

LJFgene9        -------------AGCATTGGTTCGCACCTAAGGCACC--TTCCCAATTCAGCTTCTAAC
LJFgene1        GTTCATAAGTTCCTTAATTTTCTTGAACCTTTATTTTCAGCTCCCAACAATAATGGCTTC
                              ***   * * ****       *   ******      *     *

LJFgene9        GATGACACTTTGTAGCACGTTTTCCAACTTCAA---CATTCACATATTAAAAAACAA---
LJFgene1        AATGGCATCTTCAAGCTCCTTCTGCAACCTCAAGTTTATCACCAAACCCAACAACAATGG
                 *** **  **  *** * ** * **** ****    **   ** *   ** *****

LJFgene9        ------CAAGGGTTCCTTTTCTCGTCGATTTCAACTCTCTCAGAAGCTGGATGACGATAA
LJFgene1        TAGAACCAATGCTTCTTCTCTTCCCCGTATTGTATTCTGTCAGAAGCACAACGATGACAC
                      *** * *** * *  **  **  **  * *** ********   * ** ** *

LJFgene9        TTTCATTGATAAAATCAAACGAAGGTTCTCACTGATTCTCCCTTTAATT-------TGCC
LJFgene1        CCCCACCGACCAAATCAACCGAAGGTTCTTACTTCTTCACACTCACACTCTCTCACTCCT
                   **  **  ******* ********** ***  *** * **   * *       * *

LJFgene9        ACCTCACATGAATTGTATA-----ATATATATTTATATT------TATGCTTGACCTTGA
LJFgene1        TCTTTCTATTGATTATTTATAACTATTCATACCCATCTTCTGAAATCTCTTTACATTTCA
                 * *   **  *** * **     **  ***   ** **      * *  **    ** *

LJFgene9        ATTGTTCCTATCT--TAAAGAGAGCTCATACTGGAAAGTGGAGAATTAGCAACCATTGGT
LJFgene1        ATTCTTTTTGTGTATTGAAGAGAACTCATATTGAGAAGCAGTGAAATAGCGACCATTGGT
                *** **  * * *  * ****** ****** **  ***  * *** **** *********

LJFgene9        GCCATCTTCAACTTTAGGTACACTGCTTTATTGTTTTCACAATGAGAATAGGTGACC-AA
LJFgene1        GCCATCTTCAACTTCGGGTACCCCTCCTCTGTTTTTCCTCGGTTTTTTTCTTTTTTGGAA
                **************  ***** *  * *   * *** * *  *     *   *     **
```

```
LJFgene9       AATTTAATTAATTCTTATAATATTTGAAAATATTTCGAGAAATTTTATGTAAAATTAACT
LJFgene1       AATTTAGTTTTTTCATTTTAT-TTTGAATGTA--------AATTGTATTCAAGATT---T
               ****** **  *** * * ** ******  **        **** ***  ** ***   *

LJFgene9       CTTTTTCTTTAAGATTTATGTGATTACTTCATAGCAAATTCTTGTCAAGTTTTAAGAACT
LJFgene1       GATTTTGTTGGTGGGTTTGGAGACCCTTT-----TGGATTTT-----AGTTTCAG----T
                 **** **   *  **  * **    **        *** *     ***** *     *

LJFgene9       TTTGAGTTTTGATGTGTGTTTTCTATTGGAAATATGTGGGGATATATTATTTTGGATTTT
LJFgene1       TTTGTATTGTATTG-GAAATGGGTGGTGGTTAAAAAAGAGAAAAT-TGAGTTTGGGTTTT
               ****  ** *  ** *   *   *  ***  * *   * * * ** * * ***** ****

LJFgene9       ATGCTTCTTTG---CAGAGGCAAAAAGCCAGATTATCTTGGAGTGCAGAAAAATCAACCG
LJFgene1       GTGTTTTGGTGGTGCAGTGGGAAAAAACCTGATTATCTTGGAGTGCAGAAAAACCCACCA
                ** **   **   *** ** ***** ** ********************** * ***

LJFgene9       GCATTAGCACTATGTCCGGCAACTAAGAACTGCATATCGACATCTGAAAATGTCACTAAC
LJFgene1       GCATTAGCTCTGTGTCCGGCAACTAAGAACTGCGTGTCAACCTCTGAGAATATCAGTGAT
               ******** ** ******************** * ** ** ***** *** *** * *

LJFgene9       CTCACACATTACACTCCTCCTTGGTGAAATTTCCTTCTTTATTTTTTTATTTATTAAAGT
LJFgene1       CGCACACATTATGCTCCTCCATGGTTAAAGTTCCCCTCT--TTTTCTTATTTT---AATT
               * *********  ******* **** *** ****    *  **** ******    ** *

LJFgene9       TTTAACTTTGGTTTATGATATTATCTGAATCTGAATTGGCTGCAAAGCCTGCAATTTGTT
LJFgene1       TTCACCTTCGATTTATGGGATTATATG------AATTAAATGCAATTTTTTTTACTTGTC
               ** * *** * ******  ***** **      ****   *****    *   * ****

LJFgene9       A--TTGAATATATATTGTTGTTGAGT-TTAATTTTACACTGAATAACTTTTAAAATAATT
LJFgene1       TGTTTGGATAAACAACTTAGTTAAGTATTCATCATGTAAGTGCTTATGTATAAGTTGTTT
                  *** *** * *   * *** *** ** **  *  *     * *  * ***  *  **

LJFgene9       --ATTATAAAAATCAATAAA-ATTATTATCTATATAAAAAATATGATTAGATGATAAAGT
LJFgene1       CTATAATAAATAAAAATGAGGAGGGAAAGTTATTCCAAAAATCACTTTAAAAGAGGTACT
                 ** ***** *  *** *  *     *  ***   ******    *** * **   * *

LJFgene9       AAAACTTTTGTATGGTATTAATACATGAATTATTTTTTGTCCAGAATTTTTCAGATTGTC
LJFgene1       CACTTTATTTTA-ATTATTGATTTTTTTAAAATTTAATGATTAAGATTAATTAT-TTATT
                *   * ** **   **** **   *  *  ****  **   *  ***  * *  ** *

LJFgene9       ATAAAT-TTGCATTTTGATAAAGTGTATAATT-------CACTCAATTTCAAAATAACTG
LJFgene1       ATAATTATTAGATTCTAAAAAAATAAATAATAAGGATAATAAATAACTCTAAAAAAATTA
               **** * **  *** * * *** *  *****         *   ** *  **** ** *

LJFgene9       ACTTTTTAACA--CCAAT--TTGGCAAAATGTTGCTTATGGAACAATTTAATTAATTATC
LJFgene1       TTCTTAGAGCAAGTTGATGTTTTCCTAAAAAAAATATATAAATTGTTTATATAAGTCGTA
                  **  * **     **  **  * ***       ***  *    **  ** * *  *

LJFgene9       ACATTTT--AAAACCATTG--AAAGTTCAATTGGTGTTGCCGTGTTGGGCGCATTGAGCC
LJFgene1       AGCTTTTCGGAAATTATTCTTGAAATCTTATTGAAATAATCTGAAAACAACTTTTTTTTA
               *  ****   ***  ***    ** *   ****   *   *            **

LJFgene9       CTTAACTTAGGAA--ATGTAT---TATCTCATGTTGA-ATTTGCTTTTTTTTTTTTTTTT
LJFgene1       CATGATTTGAAAACAACTTATAGACATATCATAATCACATATCATTAAGTTATTTTATTA
               * * * **   **  *  ***    ** ****  * * ** *  **   ** **** **

LJFgene9       ATTTCTCTTCA-GCTAAGTTCCGAGCATT-----GAGCCTTTAACTTAGGAAATGTATTA
LJFgene1       AGTCATTTTCATAATTTATTTCAAACACTTACATAAATACTTATAAGAGAAAATAAAATA
               * *  * ****   *   ** * * ** *      *    ***    ** ****  * **

LJFgene9       TCAC-ATGTTGGGTATAGATTATTAATTAAGAGAAC--ATTGATTATTCTACAGGAACTA
LJFgene1       AAATTAATTTCTTTATAAGCTATAAAATTAGTTAATGTATAAGTCAATAGGTAGAAGCTC
                 *  *  **   ****   *** ** * **  **   **   * * *    ** * **

LJFgene9       CA--ATCCTGAAGGTAGGAAAGATCATGTGA------GCAAAGAGGCAATGGAGGAACTG
LJFgene1       TCTCGTATTAACTCTTCAAAAGTTTATTTTTTAACTTATATATAAGCTAAATTTAACTTA
                    *  * *   *   **** * ** *          * * * ** *      *  *

LJFgene9       ATAGATGTGGTAATTTTAATTAGGATCATGTTAAGTCTT-----AAGCTACTTAGTTAAA
LJFgene1       AAAGAGAAACTTATTTCATTTTTTCTCTTCCTTTCTTTTCTAGTAAATGTTTTTATAAAA
```

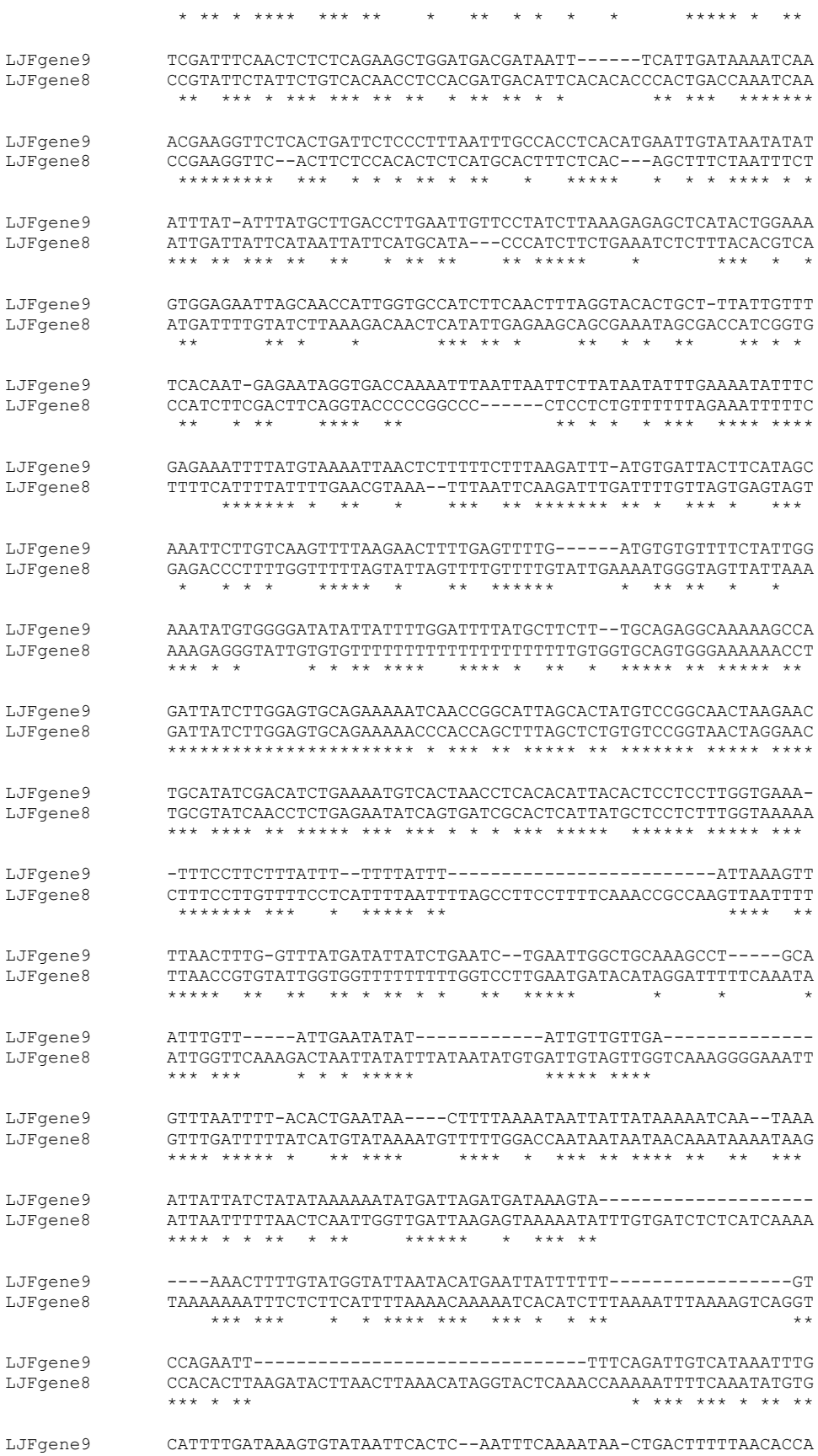
```
                * ***     * **** * **    ** *  *   * **     **     **  * ***

LJFgene9        GAATCAATAAACAATTTTTT---TAGGAAATCACGTGATTAATTACATTAGA-------A
LJFgene1        GTTTACCCAAACAGAACTTTACACAAGTACTTATAAGATAAGTCTAATTAAGCTTTTTCA
                *  *    *****    ***    * * * * *   *** * *   ****         *

LJFgene9        ATCACAGTTACACTGGCAATACTGTTATATAAAAACTAT-----TTATTTTAAATTGGTT
LJFgene1        AACATGCTCAAAGTTAAAGTGTTTCCATAATGTTTTTTTGGGCCTTAATTTTCGTCAAGT
                * **   * * * *   * *  *   ***       * *     *** ***   *    *

LJFgene9        GACTAGT-ATCTTAAGAACACTTATTAAAAAAG-TGTATTGAGTTTTTGTATCTCAATAG
LJFgene1        AACTTGTGATGTCAAAATTGCGTGTTGAGGAGGGTCTGGTCAACTAGATTTAGGAGATAA
                 *** ** ** * ** *   * * ** *  * * * *  * *  *    *      ***

LJFgene9        GAGTATAATTAAGACT-ATTTAT--AGTTGGTTTGTT---ATAGCTACTGATGA-GTTTT
LJFgene1        TTAAGCAGTCAGGGTTTATTTATTTATTGGATTTAATCCAAAAGGAACCTATGACATATT
                      * * * *  * ******  * * * ***  *   * **  **  ****  * **

LJFgene9        CAAATCATTAAAGTTACAAAAATCAGTTAC---------------TATTGTTTTTTATTT
LJFgene1        TGTGAGAATCACGTTGCAATAGATAATGGTGCACTTGGAACAATTTATCACGTTTTAAAC
                      * * * *** *** *   * *                  ***    *****

LJFgene9        TACA-GATAGAATCGACAATACTACCAGAAAATTTTACACCAAGGATTGTAGAAAGAACA
LJFgene1        CACGTGAAAGTAGTGCAACCGATATTGGACTATTGACTGTTGAACATTGTTGACTTGACA
                 **  ** ** *  *  *    **   **  ***        *  ***** **    ***

LJFgene9        GAAGA--TTATCT-TAGATTGGA-ATACCAAAGTGTATACAAGC--CACAAATTTTAACT
LJFgene1        TAAGAAATGAACTATTGGTCACACACGCATTGGCCGGTGAAAGTAGTGCAATCTTTACTT
                 ****  * * ** * * *   * *  *    *    *  ***     ***  ****  *

LJFgene9        TCAATGTCACCAATATCATTGTATGCAGAAAAAATGA-----ATAGTAA-CTTTTTACTA
LJFgene1        TTTCTTTTTTTGAAATTTTTTTTTTCACTACCTTTGGTCCTTGTATTGAGCATGGTCCCA
                *   * *     * **  ** * * **  *    **       ** * * * *  * * *

LJFgene9        TTAGACTGAAAAGCCTGC-ATCAAGCATTGAAGGA--ATGGAGTTTCTTTGATC-ATTGG
LJFgene1        CCAAAATCCAAACTTTATTATTGGACATAGATGAATCATGATGTCACTTTGTTTTAATAT
                  * * *  ***   *   **    *** ** * *  ***  **  ***** *  * *

LJFgene9        ACTGCCCATCTCATAGTAACTC--ATCTTGAAG------------CAATTAATGCAGTAA
LJFgene1        TTTCATTCTTTTTCAGGAACTATAATCCTGAAGGTAGGAAAAAACCTGTGAGCAGGGAAG
                  *     * *   ** ****   *** *****            *  * *     * *

LJFgene9        AAACTA-------ACTCATCGTGAAAAGTTCATTCTCTGCTTTATTTAAATTTTTA-CAG
LJFgene1        AAGCAATGGAGGAACTTATAGACGTGGTAATAAATGCAACTGAACTGAAATCTTGAGTTA
                ** * *       *** ** *          *    *  **  * * **** ** *

LJFgene9        CAAGTAGATTGGAATGCATGGTTTTTGCCAT-GTTTTATACTTGACAAAGATAATGCAAA
LJFgene1        TCACTGGATTGAAATTTTCTTTTCCTTCCCTCATTTTATCAACATTGATTATAATTTATA
                  * * ***** ***      **  * ** *  ******        *  *****  * *

LJFgene9        CTATAAACACAAGCAGCGG-AGCTGGTATGCACTTGGTT---TGAGATAGGATA-ATGAT
LJFgene1        AAATTTATGAAAGGAGTGATAGTGAATATACACTTTGTAACACTATTTCTAATACACTTT
                  **  *   *** ** *  **    *** ***** **      *  *   *** *   *

LJFgene9        ATATTG------------------------------------------------------
LJFgene1        CTATTATCGGTTAAAATTTATTGAAAACAGTGTTGTTAAATGGCGGCCATGGCGGCGCCA
                 ****

LJFgene9        ------------------------------------------------------------
LJFgene1        TGGCGGAGTTGCGTAACGGTTTTCTGAAAAAACGCCACCGAATAACGGTGGCGTGGCGGA


LJFgene9        ------------------------------------------------------------
LJFgene1        TTAAAGATGGCGGCGCCATGGCGGCCGCCATAGCCATGGCGGCCATGGCGGATGTGGCGG


LJFgene9        ------------------------------------------------------------
LJFgene1        GGAGGCGGAAAATGGCAGAATTTTTTTTTTTGTCCGCGGTAGGAGTTGGGCTGACCCGAT


LJFgene9        ------------------------------------------------------------
```

```
LJFgene1       CCAACCCTACCCGAAAACTTAATGAAAACCATGACCCCCCCCTACCTTTCAGAACGCTGCT

LJFgene9       ------------------------------------------------------------
LJFgene1       GCAACCCTCGAAGCTTCAACATAGCGCGACCTCGGAACCAGCCACGACCCCTGCAACCAG

LJFgene9       ------------------------------------------------------------
LJFgene1       CCTCGGAACCAGCCGCGCTTGACCCGCTGCACACAGCAGCCAGCAGCGACGACCCATGGC

LJFgene9       ------------------------------------------------------------
LJFgene1       GTGAACGGCGGCGACGAGCACGGCGTGAACAACGGCGACGCGAACGGAGAAGACGACCCA

LJFgene9       ------------------------------------------------------------
LJFgene1       GAACCGCGACCTCGACTTATACGGGTGGGTGGGCCAAAGCCTTTTTTTTGTTTTTGCTGC

LJFgene9       ------------------------------------------------------------
LJFgene1       TTGACACCCCTTTTTTATGTCTGCAGCTTTTTTTCCGTTTTTCTATTTGACACCCCCTTT

LJFgene9       ------------------------------------------------------------
LJFgene1       TACTTTCGACAGTCCCATTTTTAATTTTTTTTTCTATTTGACACCACAATTTTTTTTTCT

LJFgene9       ------------------------------------------------------------
LJFgene1       GTTCAGTCCTCTTTTAAATGGCTGAACCATCATCCTCTTTAATGAGTTTATTTGGTGTTG

LJFgene9       ------------------------------------------------------------
LJFgene1       GTACTTGATTATTGTATGAACTCATGAAACTTTTTTAGTTTATTTGAATGCAATCCTTTG

LJFgene9       ------------------------------------------------------------
LJFgene1       TTTTTTTCAATTTCAATGAGTTTATATATATGTTTTTTTTTTTTTTTGGTCCGCCATGAC

LJFgene9       ------------------------------------------------------------
LJFgene1       TTCCGCCATTTTCCGCTACGCCATCCGCCATATTTTTATGGCGGATTTTTTACTTTCCGC

LJFgene9       ------------------------------------------------------------
LJFgene1       CATGAACCGCCATCCGCCATTAACAACATTGATTGAAAATTATGAAGTCATGAGAGGAGC

LJFgene9       ------------------------------------------------------------
LJFgene1       TCATTGAATAAAGAGTGAGAACTTACATGATTTTGTAATTTCTAATAATTTTTTACTATT

LJFgene9       ------------------------------------------------------------
LJFgene1       AATAAAGAGTGTATTTAAATGGGTATGTTTTTGAACATTTTTCTAATTTCTAATCGATTT

LJFgene9       ------------------------------------------------------------
LJFgene1       GTAATAATAGCTGGTTATTTGCACTTTCCCAATCATTGAAGTAAAGTTAAACCAGTTCCG

LJFgene9       ------------------------------------------------------------
LJFgene1       GATAATTTTACAGATAGAATCAACAACACCAGACAAATTTTCACCACGGATAGTTGAAAG

LJFgene9       ------------------------------------------------------------
LJFgene1       GAAAGAAGACTATATTCGTGTGGAGTACCAAAGCTCAATCTTGGGGGTAAGTGTAACTTA

LJFgene9       ------------------------------------------------------------
LJFgene1       CATCTAAGGAAACTCATCATGAAGAAAAATTATCCTTTTATACATTTTAGATGATATCAA
```

```
LJFgene9       ------------------------------------------------------------
LJFgene1       GATTCAAGAACCATCTTTAATTTCCTTCTCCCTTTTTTCTGTCATGTGCTAACAGTTTGT

LJFgene9       ------------------------------------------------------------
LJFgene1       GGATGATGTTGAGTTCTGGTTTCCTCCGGGTAAGGGTTCTACTGTGGAGTATCGTTCTGC

LJFgene9       ------------------------------------------------------------
LJFgene1       ATCTCGGTTGGGAAACTTTGATTTTGATGTGAACAGAAAAAGAATAAAGGTATGATTCCA

LJFgene9       ------------------------------------------------------------
LJFgene1       TAATTCATATGTGCTTTCTCTATAGTTAGATAAAGAAATTCTTGGTTCCAGGGTAAAACT

LJFgene9       ------------------------------------------------------------
LJFgene1       CCCCTTTCCTTCATGTCATGTGAAGCATTTTTTACTCAAGTAGATCCACTAAATTTGAGT

LJFgene9       ------------------------------------------------------------
LJFgene1       CTCAAATGTTTTAACTTTATTCTAAATGTTTTAACTTTATTTGAGTCTCAAATGTTTTAA

LJFgene9       ------------------------------------------------------------
LJFgene1       CTGAAGGTAAATTTGGTTAACTATGATCAGAAATACATTAACAAGAGTTGAAGAATATTT

LJFgene9       ------------------------------------------------------------
LJFgene1       TTTATGTACAATAAAGAGAGTATTTGCTCGAGAGAATGTAAATCCTTTTCTAAATATTTT

LJFgene9       ------------------------------------------------------------
LJFgene1       TGTGATGAAAAATAATGGTTGCTGGCAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATG

LJFgene9       ------------------------------------------------------------
LJFgene1       GGCATCTCAAGACACCATATGATGAAAAAACTTAGGCAGAATTCACATCAGCATCTAAGA

LJFgene9       ------------------------------------------------------------
LJFgene1       AAATATTGTTTCATATACATTGTAACCTTGTATACTTTTGTATTAGATACAAAATCTCAC

LJFgene9       ------------------------------------------------------------
LJFgene1       AAGATCATTGAAAGCAAACTCTTCATGATTATTGGAATTGTAGAAATGATTGAGAACAGT

LJFgene9       ------------------------------------------------------------
LJFgene1       ACTTCAAACTCTCGGGGGAAGGAATGAAATGAAGATGTTACCCATATCTTTTTGAACTTC

LJFgene9       ------------------------------------------------------------
LJFgene1       ATTAATTGGTCCACTTATTTACTCTTTCGCTGAGTTCAATCTAACAATGTAGCATTCTTG

LJFgene9       ----------------------
LJFgene1       TTTCAAGAATTTTATGTTGTGT
```

LJFgene9 vs. LJFgene8

```
LJFgene9       ----------------------------------------------------------AGCATT
LJFgene8       ATTGACTTGAAAGATTCTTGTAGCAATTTGCAGCAGTTTTATATAGATAGTAACATTCTT
                                                                        *   **

LJFgene9       GGTTCGCACCTAAGGCACCTTCCCAATTCAGCTTCTAACGATGACACTTTGTAGCACGTT
LJFgene8       AAATTACGTTTATAAGTTCCTTCATTTTCAGCTCCGAACAATGGCTTCTTCGTTCTCCTT
                  *  *   **      * * *   ******* * *** *** *   **    * * **

LJFgene9       TTCCAACTTCAACATTCACAT---ATTAAAAAACAACAAGGG------TTCCTTTTCTCG
LJFgene8       CTGCACCCTCAAGTTTCGCACCAAACCCAACGATAGTAGAAGCAGTGCTTCCTCTCTTCC
```


LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9
LJFgene8
LJFgene9

```
LJFgene8       AGATCCCACACAAATTTTTATTTATTAGAAATCTTAAAATAAATTAAATGTTTATCAGGA
                  *   * * *   * * *** * *   *** * *******  * * * **** **  *

LJFgene9       ATTTGGCAAAATGTT-GCTTATG---GAACAATTTAATTAATTATCACATTT-TAAAACC
LJFgene8       ATTTGAATGTATTTTAGCCTATTTTCGAACAAGGTTTTATCATACAAAATTTATGGATCC
               *****     ** ** ** ***    ******  *  *    **  * **** *  * **

LJFgene9       ATTGAAAGTTCAATTGGTG-TTGCCGTG--TTGGGCGC----ATTGAGCCCTTAACTTAG
LJFgene8       AACAAAATAAAAATAAGAAATTGTCATGGCTTGGTTTCCAAAACTGAACCTCATGTTCAA
               *   ***    ***  *   *** * **  ****   *    * *** **      * *

LJFgene9       GAA-ATGTATTATCTCATGTTGAATTTGCTTTTTTTTTTTTTTTTTATTTCTCTTCAGCTA
LJFgene8       GAACACGTGGAAGCACAACTGCAGCTTTGTCTTAGCTCATCATCAAGCGCCAGTTAGCAA
               *** * **   * * **  *  *  **  * **   *  *  *  *   *   * *** *

LJFgene9       AGTTCCG-AGCATTGAGCCTTTAACTTAGGAAATGTATTAT-CACATGTTGGGTATAGAT
LJFgene8       TATTTTGCGGGAATAGACAATGAGACCAAAACCTTCATCATGCACAAGTCTATTTTAGTT
                 **  *  * * *   *  * *    *  *  *  ** ** **** **    * *** *

LJFgene9       -TATTAA--TTAAGAGAACATTGATTATTCTACAGG-AACTACAATCCTGAAGGTAGGAA
LJFgene8       GCACTAGGGTTTTGGAGCCTTTCACAAGACCAAAACTAGATACGATTCATGCTAAAATAA
                 * **   **  *    * ** *  *  * * *   *  *** ** *       *  **

LJFgene9       AGATCATG----TGAGCAAAGAGGCAATGGAGGAACTGATAGATGTGGTAATTTTAATTA
LJFgene8       AGCCTAAAGCCTTTTGGAATTATTCACAGCAAAATCCGATGTTGACATCAACCATATGTA
               **   *      *  * **  *  **  * *  * * ***         **   **  **

LJFgene9       GGATCATGTTAAGTCTTAAGCT------ACTTAGTTAAAGAATCAATAAACAATTTTTTT
LJFgene8       ACATTATAATATATTACTAGCCTTTAAAATTCAATTCCACATTCAACAAAATTTATTACC
                 ** **  **  *    ***       * * * **  * * **** ***   * **

LJFgene9       AGGAAATCACGTGATTAATTACATTAGAAATCACA--GTTACACTGGCAATACTGT----
LJFgene8       CTGTTCTCTCTACACTACGCA-ATTAGAAAACAAAAGGTTGAGCGGAGAAAAATCTAGAA
                 *   ** *   * **   * ******** ** *  ***   * *  ** * * *

LJFgene9       --TATATAAAAACT------ATTTATTTTAAATTGGTTGA-CTAGTATCT-----TAAGA
LJFgene8       AGTGCTTAGAAATTCAACTAAATTTTGTCAAACAGGACGAGCCAATAAATACGATTAATG
                 *   ** *** *      * ** * * ***  **  ** * * **  *     ***

LJFgene9       ACACTTATTAAAAAAGTGTATTG--AGTTTTTGTAT-----CTCAATAGGAGTATAATTA
LJFgene8       ACAGCTACGAGAAATGTATATTTTTAGTTTAGAAACACGTCTTTAGTTATATCATAACAA
               ***  **  * *** ** ****   *****    *       * * *   *  ****  *

LJFgene9       AGACTATTTATAGTTGGTTTGTTATAGCTACTGATGAGTTTTCAAATCATTAAAGTTACA
LJFgene8       AAAATAAAAATTAAAGTTGTGTTTAAATATATGTTTTATTTAAAAAAATTAAAAATTAGG
               * * **   **    * * ****  *     ** *   ***  ***   * *** ***

LJFgene9       AAAATCAGTTACTATTGTT--TTTTATTTTACAGATA-GAATCGACA-ATACTACCAGAA
LJFgene8       AAATTGTGTTTTAATTCATGACTTTAAGTTAAAGTCACGAATTAAAAGATGGCTTTAATA
               *** *  ***   ***  *   ****  *** **  * ****  * * **      *  *

LJFgene9       AATTTTACACCAAGGATTGTAGAAAGAA--------CAGAAGAT-TATCTTAGATTGGAAT
LJFgene8       AAAGTAAATTTAAAAAATAAATAAAGAAGTCATGTTTGAAAATATACCTTTCGTTGAAAT
               **  * *    **  * *  * ******         *** ** ** ***   *** ***

LJFgene9       ACCAAAGTGT----ATACAAGCCACAAATTT--TAACTTC--------AATGTCACCAAT
LJFgene8       GTTGTGCTTTTAAAACACAACTTGCTCATTTATTGATTTTTTAAAAATAAATTTACCTAT
                      * *    * ****    *  ****  * * **         **  * *** **

LJFgene9       ATCA-----TTGTATGCAGAAAAAATGAA------TAGTAACT---TTTTACTATTAGAC
LJFgene8       TTTGGGATTTTTTTTTTATAAAAAATGCACCACTCTAGTAAGTAAACCTTTTTTTTATAT
                *       ** * *  * ******** *      ****** *     **  * *** *

LJFgene9       TGAAAAGCCTGCATCAAGC--ATTGAAGGAATGGAGTTTCT-TTGAT-CATTGGACTGCC
LJFgene8       ATAAAAAATTGCATTCAGGTAAACAACATAATTAAGTGTTTACTGATTCATTGAAACACT
                 ****   *****  **   *   *   ***  *** * *  **** ***** *   *

LJFgene9       CATCTCATAGTAACTCATCT--TGAAGCAATTAATGCAGTAAAA--ACTAACTCATCGTG
LJFgene8       TATGTATAAGTTGTTTATGTGATTGAAGAGAAAATAAAGTTAAATTATTTTCTTATAAAT
                ** *   ***   * ** *  *  *  *   ***  *** ***  * *  ** **
```

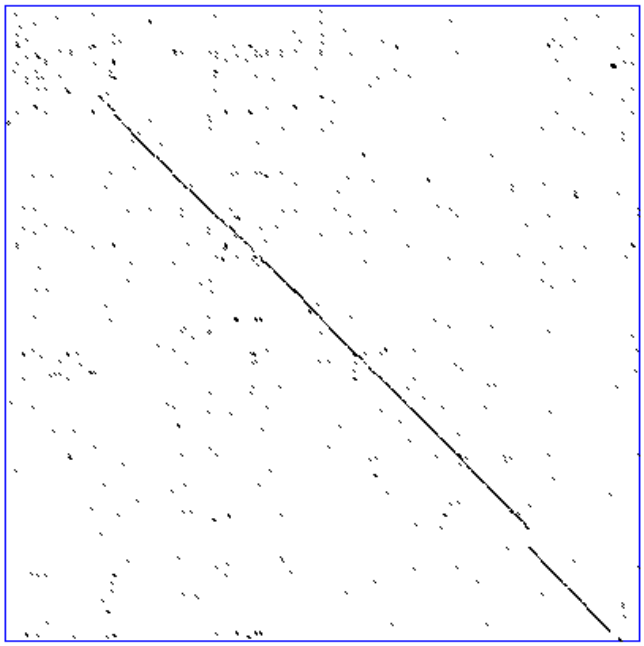
```
LJFgene9       AAAAGTTCATTCTC-TGCTTTATTTAAATTTT--TACAGCAAGTAGATTGGAATGCATGG
LJFgene8       TGTAATATGTTTTCATGAGCTATGGAAAGTTTATTGAAATAAACTGAAAATAGATTGTGG
                  * *   ** ** **   ***  *** ***  *  *  **   **    *     ***

LJFgene9       TTTTTGCCATGTTTTATACTTGACAAAGATAATGCAAACTATAAACACAAGCAGCGGAGC
LJFgene8       ATATTTCATAAATACATATCTTAAAATTATTTTAAATATTCCAAACACTTATATAAATAC
                * ** *     *  ***  * * **  **  *  * * *  ******    *      *

LJFgene9       TGGTATGCACTTG-GTTTGAGATAGGATAATGATATATTG--------------------
LJFgene8       TA-TAAGCACTTGTAATAGAAGAAAGATAAGAATGTAAAATAAATTGATTTTTTTTCCAT
               *  ** *******   * **   * *****  ** **

LJFgene9       ------------------------------------------------------------
LJFgene8       AAGTTGATTTAAGTTAAAATCAACTTATGTATCATAACACCTCAGGTTTTGAAAAAGTTA

LJFgene9       ------------------------------------------------------------
LJFgene8       AATGAGAGAGTTTTTAACAAAGTTAAGTGTATAAGTTAAATTTAAGAGAAAAACACAATT

LJFgene9       ------------------------------------------------------------
LJFgene8       TATTCTACCTTTTCTTTTCCCCATTTGTAATTGTTTATGGCCAATTTGATCCAAACAACA

LJFgene9       ------------------------------------------------------------
LJFgene8       TCTTTAGCCTAAATAAGCTCTTCCAATCACACTCTAAGTTTAAGTATGATTACTATGATG

LJFgene9       ------------------------------------------------------------
LJFgene8       TTTAGATTTCATTGTGTTGTTTTTTATGACCTTAATTTTCCGTCGAGTAACGCCATATTT

LJFgene9       ------------------------------------------------------------
LJFgene8       GAATTAAAGTTTGTAAATTTTAGATGAAATTCATTTTGCAAACTGAATTTGTAAAATCAA

LJFgene9       ------------------------------------------------------------
LJFgene8       CTCTTCCTTTACTTCCAAAATCCAAACTTTATTATTGGACAAAGATGAATCGTGGAGTTA

LJFgene9       ------------------------------------------------------------
LJFgene8       CTTTGTTTTAAATTTTTAATTTTAATTATTGTTCAGGAACTACAATCCTGAAGGTAGGAA

LJFgene9       ------------------------------------------------------------
LJFgene8       AAACCCTGTGAGCAGAGAAGAGGCAATGGAGGAACTGATAGACGTGGTAATAAATCTAGC

LJFgene9       ------------------------------------------------------------
LJFgene8       TGAAATCATAAGTTATTTTCATGAATGCATTGCAATTTCCCTTTCCTAGCCTGTGTATCA

LJFgene9       ------------------------------------------------------------
LJFgene8       ACAAATGTTATTATTTATAATAAGTTTAATTTTCATGCACTGACCGTATATAATAATTTT

LJFgene9       ------------------------------------------------------------
LJFgene8       ATATTGATATCTAATCATAAATCATCATTTAAATTATTTTAAGATAATTAATTTAAAAGT

LJFgene9       ------------------------------------------------------------
LJFgene8       TAATAAATTTACCGTATATAATGAATTATAATTGAATAATTGTATAAAAAATTTATAATG

LJFgene9       ------------------------------------------------------------
LJFgene8       TCACTGTATAATTCTTTTTCTCATTTATAATTTGGTTGTAAGTTGTAGCTTGTAATAGCA

LJFgene9       ------------------------------------------------------------
LJFgene8       GGTTATTTACCTTTTTCCAATCATTGAAGTAAAGTTAAAATCCAGCTCTGGATAATTTTA
```

```
LJFgene9       ------------------------------------------------------------
LJFgene8       TAGATAGAATCAACAACACCAGACAAATTTACACCACGAATAGTTGAAAGGAAGGAAGAC

LJFgene9       ------------------------------------------------------------
LJFgene8       TATATTCATGTGGAGTACCAAAGCTCAATCTTGGGGGTATGTGTAACTTACATCAAAAGG

LJFgene9       ------------------------------------------------------------
LJFgene8       AAACTCATCGTGGAGAAAAATAATAATTTTGTACATTTTAGATGATAATCAAGAACCATC

LJFgene9       ------------------------------------------------------------
LJFgene8       TCTAATCCCTTCTCCCCTCCTTTTTATTTTTTCTGCCATGTGCTAGCAGTTTGTGCATGA

LJFgene9       ------------------------------------------------------------
LJFgene8       TGTTGAGTTCTGGTTTCCACTGGGTAAGGGTTCTACTGTGGAGTATCGATCTGCATCTCG

LJFgene9       ------------------------------------------------------------
LJFgene8       GTTGGGGAACTTTGATTTTGATGTGAATAAGAAAAGAATAAAGGTATGTTTGTATCATTC

LJFgene9       ------------------------------------------------------------
LJFgene8       CTTTGTGCTGTCTCGGTAGTTAACATGAAGAAATGATTAAAAGATATTTTGTCCTTTAGG

LJFgene9       ------------------------------------------------------------
LJFgene8       TTTTTTGGTTATATTTAGTTTGATTTTTTATTTTTTAAAAGTTAAATTTAGTCCTTTAT
```

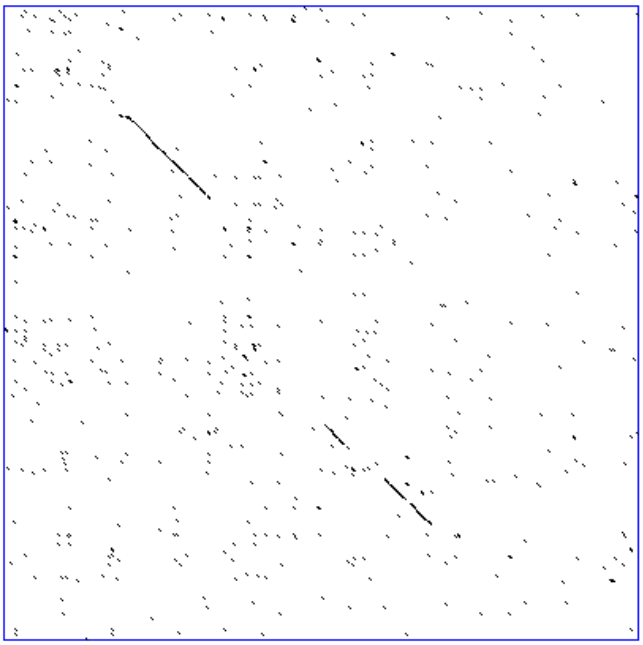
APPENDIX I.

ADDITIONAL DOT PLOT MATRICES

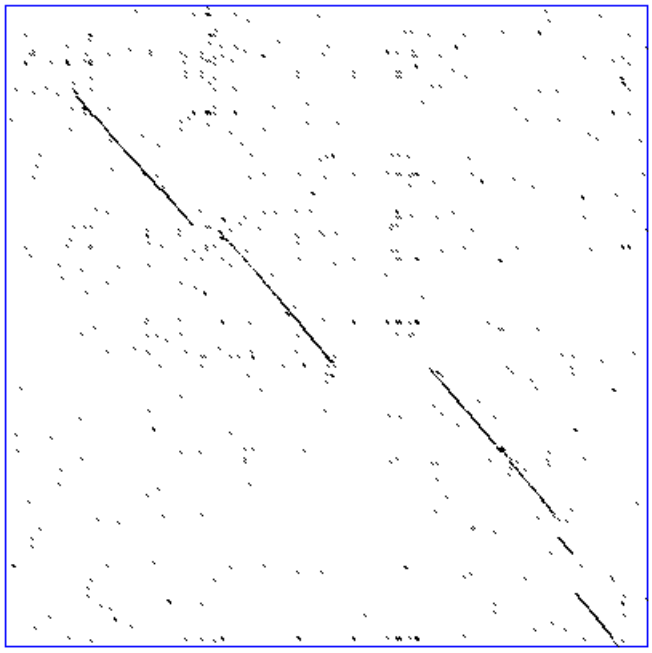
LJFgene3 plus 1000nt extension from both 5’ and 3’ gene model boundaries (x-axis) vs. LJFgene14 plus 1000nt extension from both 5’ and 3’ gene model boundaries (y-axis)



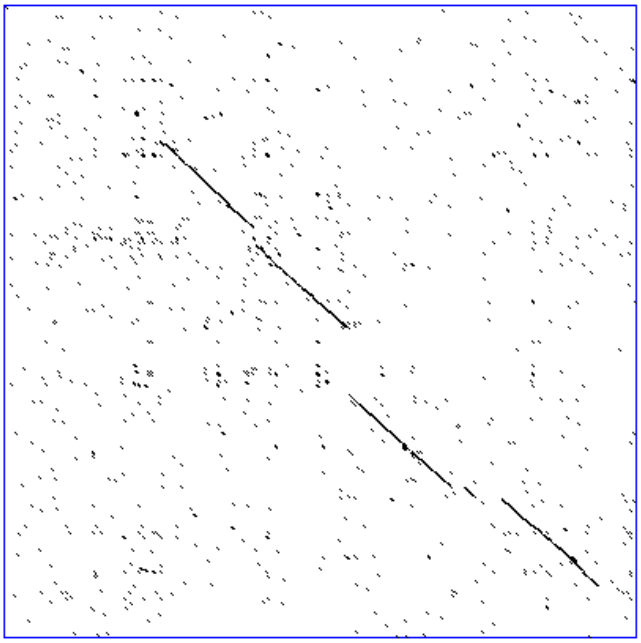
LJFgene3 plus 1000nt extension from both 5’ and 3’ gene model boundaries (x-axis) vs. LJFgene8 plus 1000nt extension from both 5’ and 3’ gene model boundaries (y-axis)



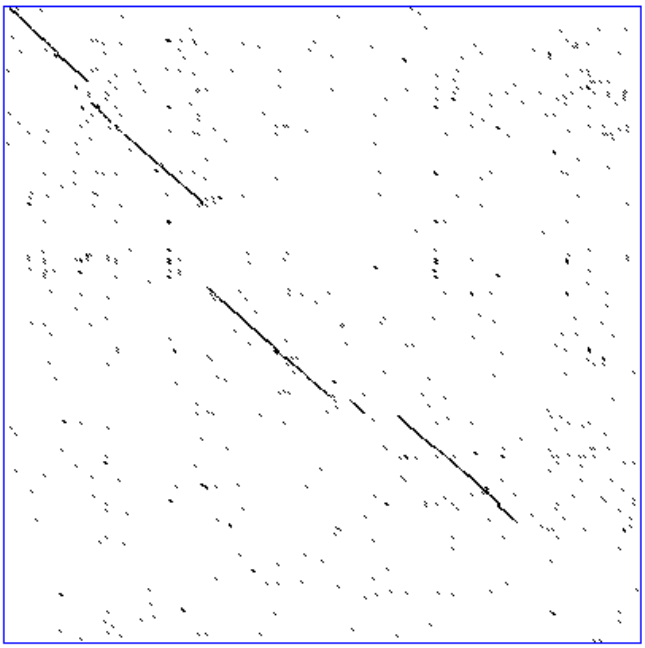
LJFgene1 plus 1000nt extension from both 5’ and 3’ gene model boundaries (x-axis) vs. LJFgene14 plus 1000nt extension from both 5’ and 3’ gene model boundaries (y-axis)



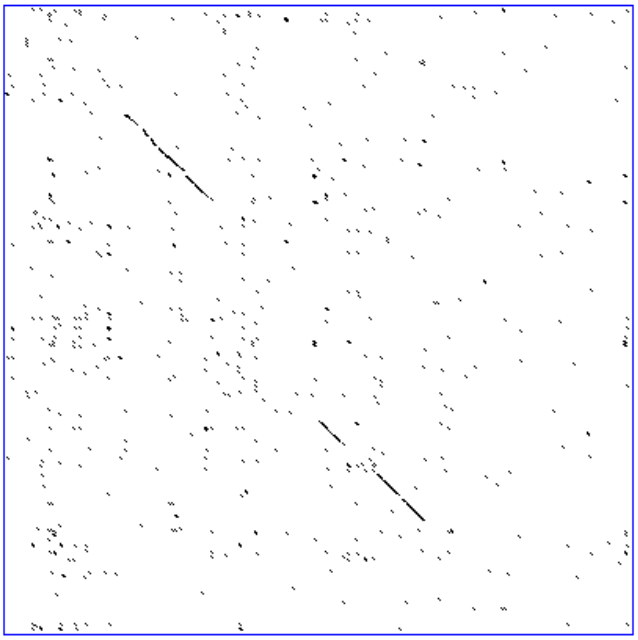
LJFgene14 plus 2200nt extension from both 5’ and 3’ gene model boundaries (x-axis) vs. LJFgene1 plus 2200nt extension from both 5’ and 3’ gene model boundaries (y-axis)



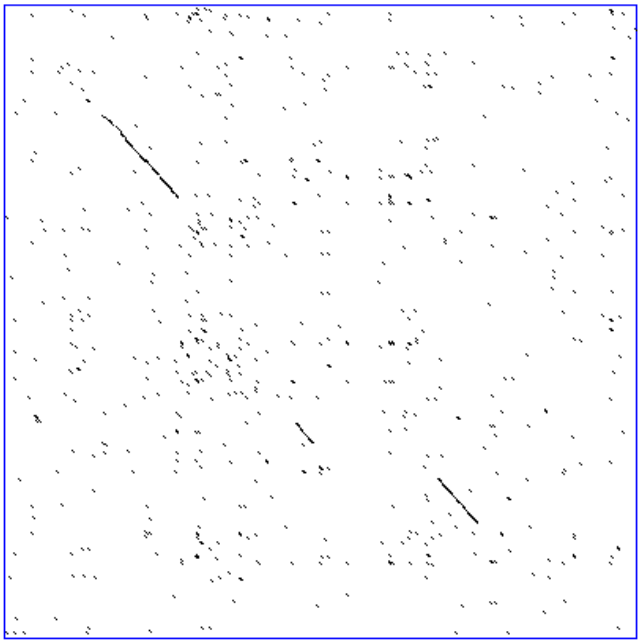
LJFgene14 plus 2200nt extension from both 5’ and 3’ gene model boundaries (x-axis) vs. LJFgene1 plus 2200nt extension from both 5’ and 3’ gene model boundaries (y-axis). Adjustment: 5’ extension removed to shift plot for 3’ similarity analysis



LJFgene14 plus 1000nt extension from both 5’ and 3’ gene model boundaries (x-axis) vs. LJFgene8 plus 1000nt extension from both 5’ and 3’ gene model boundaries (y-axis)



LJFgene1 plus 1000nt extension from both 5’ and 3’ gene model boundaries (x-axis) vs. LJFgene8 plus 1000nt extension from both 5’ and 3’ gene model boundaries (y-axis)



APPENDIX J.

DIVERSE PLANT FAMILY PEPTIDE ALIGNMENT

```
n101401_S.      ------------------------------------------------------------
Pp1s10_47V      --------------MPPILVLASSFTGSKLHIPSALQDKVSISEVSNASPGIIVCQQKED
LJFgene3        MSISSLIFSNLHFQLPTTMASMASSSSFCN-LKFITKPNN-GRR--SSLPRIVFCQKHHD
LJFgene1        ------------------MASMASSSSFCN-LKFITKPNNNGRTNASSLPRIVFCQKHND
LJFgene14       ------------------------------------------------------------
Phvul.010G      ------------------MASLIPSSSFCSYLKVCTKPSN-GRISASSFPRILCCQKHHH
LJFgene8        ---------------------MASSFSFCT-LKFRTKPND-SRSSASSLPRILFCHNLHD
Medtr4g023      ------------------MASTSSSTTFCNLKFHSTRPNN---NNVSFLPRILHMKQEEE
LJFgene9        ------------------------------------------------------------
AT3G60810.      ------------------MASMASPTTCLYHHKCRRKLVFYARRISASIPETSYDKHP--
LOC_Os03g6      ----------------------MAPPPPCCVLVVRCCSGSLPSPPPPPPPRPSPSNNLR-
Cre11.g468      ---------------------MLQAQKRVVFGQAQRRSAVA-VVRTAPVARRMVCQAANW
Vocar20006      ---------------------MQLQRSPCSRAVAQRKCALKPAMRIAAPALQVACQAGRW
Pp1s223_13      ------------------MMPIPHACAAPWQRNCHRIVASIPAAEGSSFPKIAAGSVGAQ


n101401_S.      ------------------------------------S-GTRPDYLGIQ----KNPPSLA
Pp1s10_47V      EQVSGQIRKEAVAAKISRRETILRSSGSALMLAFFNFA-GERPDYLGVQ----SNPPSLA
LJFgene3        ST---------PTDQINRRELILRSSEIATIGAILNFG-GKKPDYLGVQ----KNPPALA
LJFgene1        DT---------PTDQINRRELILRSSEIATIGAIFNFG-GKKPDYLGVQ----KNPPALA
LJFgene14       -------------------------MGGLGFVFWWCSGKKPDYLGVQ----KNPPALA
Phvul.010G      DV---------PTDQINRRELILRSSEIATIGAIFNLS-GKKPDYLGVQ----KNPPALA
LJFgene8        DIHT-------PTDQINRRQLILRSSEIATIGAIFDFS-GKKPDYLGVQ----KNPPALA
Medtr4g023      DN---------NTNQINRRQLILRSSEIATIGAIFNFS-GKKPEYLGVQ----KNSSALA
LJFgene9        ----------------------------ATIGAIFNFR-GKKPDYLGVQ----KNQPALA
AT3G60810.      -------------KLIGRRDIILRSSELAMIGAIFQLS-GKKPDYLGVQ----KNE-RLA
LOC_Os03g6      ---------------IARREFVLRSSELATLAAIFHLS-GTKPRYLGVQ----KSPPSLA
Cre11.g468      K--------GDSEEKLQPKEIVLRSVNVMVLGALLSIGAAPRPGNLGIIDY-GAGVQTLN
Vocar20006      QA-------GENDEKLQPKEVILRSVNMAVVAALLTIGAAPRPSNLGIQDY-GAGIQTLS
Pp1s223_13      ESNQKIN--ILLAGVVSAAVCVNVSLQGCAVAASTNVPNNGRTWFPWEQTTPEQFEPEQG
                                                        :.

n101401_S.      LCPPTPNCISTSE-EANDPSHYVPPWTYNPEDGRGRKN-PATREQAMKELLAVISSTK-P
Pp1s10_47V      LCPPTPGCISTSE-ELNDPTHYVPPWTYNPPDGRGRKN-PASKEKAMAELIDAIKSTK-P
LJFgene3        LCPATKNCVSTSE-NISDRTHYAPPWNYNP---EGRKK-PVNREEAMEELIDVIESTT-P
LJFgene1        LCPATKNCVSTSE-NISDRTHYAPPWNYNP---EGRKK-PVSREEAMEELIDVIESTT-P
LJFgene14       LCPPTKNCVSTSE-NISDRTHYAPPWNYNP---EGRKK-PVSREEAMEELIDVIESTT-P
Phvul.010G      LCPATKNCVSTSE-NISDRTHYAPPWNYNP---EGRKN-PVSKEEAMEELIDVIESTT-P
LJFgene8        LCPVTRNCVSTSE-NISDRTHYAPLWNYNP---EGRKN-PVSREEAMEELIDVIESTT-P
Medtr4g023      LCPATKNCVSTSE-NVNDLTHYAPPWNYNP---EGRKS-PVSREEAVEELIEVIELTR-P
LJFgene9        LCPATKNCISTSE-NVTNLTHYTPPWNYNP---EGRKD-HVSK-EAMEELIDVIESTILP
AT3G60810.      LCPATNNCISTSE-NISDRVHYAPPWNYN----GGRKT-PVNRQVAMKELLNVIKSVK-P
LOC_Os03g6      LCPATNNCVSTCE-DITDSIHYAPPWNYNPK--DGRRAKPITKHEAINQLIQVVTQTK-P
Cre11.g468      LCPPSPNCIATSE-EGNDRTHYAPPLTYNPEDGRGKKG-PASQEKAMGELVEAVKKLK-P
Vocar20006      LCPSTPNCIATSE-EGNDREHYAPPLTYNPQDGRGRKN-PATQEQAMAELVSVVKTLQ-P
Pp1s223_13      TCSTCIGVVDDTLGSCSATTNCVSSFDDRPSFFTAPWEFPGSLRAAVSNLQEALEGSG--
                 *.   . :     . .   : ..    .     .      .   *: :*  .:

n101401_S.      ENYTPNIVKNTDDYVYVEYQSPYLGFVDDVEFWFP-PGNRSLVEYRSASRVGSS-DFDAN
Pp1s10_47V      DNFTPRIVKQTDDYVYVEYSSPLVGFVDDVEFWFP-PGNRSLVEYRSASRRDAI-DFGFN
LJFgene3        DKFSPRIVERKEDYIRVEYQSSILGFVDDVEFWFP-PGKGSTVEYRSASRLGNF-DFDVN
LJFgene1        DKFSPRIVERKEDYIRVEYQSSILGFVDDVEFWFP-PGKGSTVEYRSASRLGNF-DFDVN
LJFgene14       DKFSPRIVERKEDYIRVEYQSSILGFVDDVEFWFP-PGKGSTVEYRSASRLGNF-DFDVN
Phvul.010G      DKFTPRIVERKEDYIRVEYQSSILGFVDDVEFWFP-PNKGSTVEYRSASRLGNF-DFDLN
LJFgene8        DKFTPRIVERKEDYIHVEYQSSILGFVHDVEFWFP-LGKGSTVEYRSASRLGNF-DFDVN
Medtr4g023      DKFTPKIVERKEDYVRVEYRSSILGFVDDVEFWFP-PGKGSIVEYRSASRLGNF-DFDVN
LJFgene9        ENFTPRIVERTEDYLRLEYQS-----VYKPQILTS-MSPISLYAEKMNSNFLLL-D----
AT3G60810.      DKFTPRIVEKKDDYVHVEYESPILGLVDDVEFLFT-PGKNSKVEYRSASRKGNF-DFDVN
LOC_Os03g6      DNFTPRLVEKTDDYVRVEYESPIFGFVDDVEFWFP-PGNKSIVQYRSASRSGFI-DFNAN
Cre11.g468      DGFTPKIIKQTDDYLYVEYESPLMGFIDDVEFWFK-PGPGSRVEYRSASRVGES-DGNIN
Vocar20006      DGFTPKIIKQTPNYLYVEYESPIMGFIDDVEFWFK-PGPGARVEYRSASRVGES-DGNIN
Pp1s223_13      ----AFIKEKSDRYIYAVFKGD-DGVSDDVEFLFSDPSVDATVNVRSASRAKDYKDSGRN
                    . : :..  *:   : .       . ::     .  :    :  *.     *

n101401_S.      RKRIRALRKALEKKGWQSIGF---------------------------------
```

```
Pp1s10_47V      RKRIKALRQALERYGWESIGF----------------------------------
LJFgene3        RKRIKALRQELEKKGWASQDTI---------------------------------
LJFgene1        RKRIKALRQELEKKGWASQDTI---------------------------------
LJFgene14       RKRIKALRQELEKKGWTSQDTI---------------------------------
Phvul.010G      RKRIKALRQELEKKGWASEDTI---------------------------------
LJFgene8        KKRIKVC---LYHSFVLSR------------------------------------
Medtr4g023      RKRIKALRQELEKKGWASQDTTI--------------------------------
LJFgene9        ----------------------------------------------------------
AT3G60810.      RKRIKALRQELEKKGWVSENSF---------------------------------
LOC_Os03g6      KKRVKALRLALENKGWASESTI---------------------------------
Cre11.g468      RKRIKAIRQELETKGWRSTGF----------------------------------
Vocar20006      RKRIRAIRQELEKEGWRSTGF----------------------------------
Pp1s223_13      RKRLEALRMELGWEQVPILRNRQRRLFFIESPWDTFGPEPPPTIDYKNGIDFVPD
```

APPENDIX K.

DIVERSE PLANT FAMILY CODON ALIGNMENT

```
n101401_S.    ------------------------------------------------------------
Pp1s10_47V    ----------------------------------------ATGCCTCCCATCTTGGTT
LJFgene3      ATGTCCATAAGTTCCTTAATTTTCTCGAACCTTCATTTTCAGCTCCCAACAACAATGGCT
LJFgene1      ------------------------------------------------------ATGGCT
LJFgene14     ------------------------------------------------------------
Phvul.010G    ------------------------------------------------------ATGGCT
LJFgene8      ------------------------------------------------------------
Medtr4g023    ------------------------------------------------------ATGGCA
LJFgene9      ------------------------------------------------------------
AT3G60810.    ------------------------------------------------------ATGGCT
LOC_Os03g6    ------------------------------------------------------------
Cre11.g468    ------------------------------------------------------------
Vocar20006    ------------------------------------------------------------
Pp1s223_13    ------------------------------------------------------ATGATG

n101401_S.    ------------------------------------------------------------
Pp1s10_47V    CTTGCTTCGTCCTTCACTGGTTCCAAACTCCACATACCATCCGCACTTCAGGATAAGGTT
LJFgene3      TCAATGGCATCTTCAAGCTCCTTCTGCAAC---CTCAAGTTCATCACCAAACCCAACAAT
LJFgene1      TCAATGGCATCTTCAAGCTCCTTCTGCAAC---CTCAAGTTTATCACCAAACCCAACAAC
LJFgene14     ------------------------------------------------------------
Phvul.010G    TCACTGATACCTTCAAGCTCTTTCTGTAGCTATCTCAAGGTTTGCACCAAACCCAGCAAT
LJFgene8      ---ATGGCTTCTTCGTTCTCCTTCTGCACC---CTCAAGTTTCGCACCAAACCCAACGAT
Medtr4g023    TCAACATCATCTTCAACCACATTCTGCAACCTCAAGTTTCACTCCACACGACCCAACAAC
LJFgene9      ------------------------------------------------------------
AT3G60810.    TCCATGGCGTCACCGACCACTTGCTTGTACCACCACAAATGCCGGAGAAAACTTGTTTTC
LOC_Os03g6    ------ATGGCGCCGCCGCCGCCGTGTTGTGTGCTAGTGGTACGCTGCTGTTCCGGATCC
Cre11.g468    ---ATGCTGCAGGCCCAGAAGCGTGTTGTGTTTGGCCAGGCCCAGCGCCGCAGCGCTGTT
Vocar20006    ---ATGCAGCTTCAGAGGTCTCCGTGCAGCAGGGCCGTCGCCCAGCGCAAGTGCGCATTA
Pp1s223_13    CCAATTCCGCATGCATGTGCTGCCCCTTGGCAGCGGAATTGTCATCGCATTGTAGCGTCC

n101401_S.    ------------------------------------------------------------
Pp1s10_47V    TCTATTTCTGAGGTTTCCAATGCTTCTCCTGGTATTATCGTATGTCAACAAAAAGAAGAT
LJFgene3      ---GGTAGAAGA------AGCTCTCTTCCCCGTATTGTATTCTGTCAGAAGCACCACGAT
LJFgene1      AATGGTAGAACCAATGCTTCTTCTCTTCCCCGTATTGTATTCTGTCAGAAGCACAACGAT
LJFgene14     ------------------------------------------------------------
Phvul.010G    ---GGTAGAATCAGTGCTTCCTCTTTTCCTCGCATTCTCTGCTGTCAGAAGCACCACCAT
LJFgene8      ---AGTAGAAGCAGTGCTTCCTCTCTTCCCCGTATTCTATTCTGTCACAACCTCCACGAT
Medtr4g023    ---------AACAATGTTTCCTTTCTTCCTCGCATTCTTCACATGAAACAGGAGGAGGAG
LJFgene9      ------------------------------------------------------------
AT3G60810.    TATGCCCGTCGTATCAGTGCCTCGATTCCAGAGACTTCATATGACAAACATCCG------
LOC_Os03g6    CTCCCTTCGCCGCCGCCGCCGCCGCCCAGACCCTCTCCCTCCAACAACCTCAGA---
Cre11.g468    GCC---GTTGTTCGCACTGCCCCTGTGGCTCGCCGCATGGTCTGCCAGGCGGCCAACTGG
Vocar20006    AAGCCCGCGATGCGCATTGCTGCACCGGCGCTGCAGGTGGCGTGCCAGGCTGGGCGTTGG
Pp1s223_13    ATTCCAGCCGCAGAGGGTTCAAGCTTCCCGAAGATTGCTGCAGGTAGTGTAGGAGCACAG

n101401_S.    ------------------------------------------------------------
Pp1s10_47V    GAACAGGTATCAGGGCAAATACGGAAAGAGGCTGTGGCTGCAAAGATATCACGAAGAGAA
LJFgene3      AGCACA--------------------------CCCACCGACCAAATCAACCGAAGAGAA
LJFgene1      GACACC--------------------------CCCACCGACCAAATCAACCGAAGAGAA
LJFgene14     ------------------------------------------------------------
Phvul.010G    GATGTT--------------------------CCCACCGACCAAATCAACCGAAGAGAA
LJFgene8      GACATTCACACA--------------------CCCACTGACCAAATCAACCGAAGACAA
Medtr4g023    GACAAC--------------------------AACACCAACCAAATCAATCGAAGACAA
LJFgene9      ------------------------------------------------------------
AT3G60810.    ----------------------------------------AAGCTAATTGGTCGAAGAGAT
LOC_Os03g6    ----------------------------------------ATCGCCAGAAGAGAG
Cre11.g468    AAG----------------------GGCGACAGCGAGGAGAAGCTCCAGCCCAAGGAG
Vocar20006    CAGGCG-------------------GGCGAGAACGATGAGAAGCTCCAACCCAAGGAA
Pp1s223_13    GAAAGTAATCAGAAGATTAAT------ATCCTGCTGGCTGGTGTGGTGAGTGCCGCCGTG

n101401_S.    ----------------------------------------------------AGT---GGA
Pp1s10_47V    ACAATTTTGAGATCTAGTGGTTCCGCGCTGATGTTGGCCTTCTTCAACTTCGCC---GGC
LJFgene3      CTCATATTGAGAAGCAGCGAAATAGCGACCATTGGTGCCATCTTGAACTTCGGT---GGG
LJFgene1      CTCATATTGAGAAGCAGTGAAATAGCGACCATTGGTGCCATCTTCAACTTCGGT---GGG
LJFgene14     --------------------ATGGGTGGTTTGGGTTTTGTGTTTTGGTGGTGCAGTGGG
```

```
Phvul.010G    CTCATATTGAGAAGCAGTGAAATAGCGACCATTGGTGCCATCTTCAACCTCAGT---GGA
LJFgene8      CTCATATTGAGAAGCAGCGAAATAGCGACCATCGGTGCCATCTTCGACTTCAGT---GGG
Medtr4g023    CTCATATTGAGGAGCAGTGAAATAGCAACGATTGGTGCCATCTTCAATTTCAGT---GGG
LJFgene9      ----------------------GCAACCATTGGTGCCATCTTCAACTTTAGA---GGC
AT3G60810.    ATCATTCTTAGGAGCAGTGAATTAGCTATGATCGGAGCCATATTCCAGCTTAGT---GGG
LOC_Os03g6    TTTGTGCTGAGGAGCAGTGAGCTCGCCACGCTCGCTGCCATCTTCCACTTGAGC---GGC
Cre11.g468    ATTGTGCTCCGCTCCGTGAACGTTATGGTTCTGGGCGCGCTGCTGTCCATCGGCGCGGCC
Vocar20006    GTCATCCTCCGGTCCGTGAACATGGCCGTCGTGGCCGCGCTCCTGACAATCGGAGCTGCT
Pp1s223_13    TGTGTCAACGTGAGCCTCCAGGGGTGCGCTGTGGCAGCCAGCACTAATGTGCCCAACAAT

n101401_S.    ACGCGACCTGACTATCTTGGAATACAA----------AAAAATCCGCCGTCCCTTGCG
Pp1s10_47V    GAAAGACCAGACTACCTAGGTGTACAG----------TCGAACCCACCGTCTCTTGCA
LJFgene3      AAAAAACCTGATTATCTTGGAGTGCAG----------AAAAACCCACCAGCATTAGCT
LJFgene1      AAAAAACCTGATTATCTTGGAGTGCAG----------AAAAACCCACCAGCATTAGCT
LJFgene14     AAAAAACCTGATTATCTTGGAGTGCAG----------AAAAACCCACCAGCATTAGCT
Phvul.010G    AAAAAACCAGATTATCTTGGAGTGCAG----------AAAAACCCACCAGCATTAGCT
LJFgene8      AAAAAACCTGATTATCTTGGAGTGCAG----------AAAAACCCACCAGCTTTAGCT
Medtr4g023    AAGAAGCCTGAATATCTTGGAGTTCAG----------AAAAACTCATCAGCATTGGCT
LJFgene9      AAAAAGCCAGATTATCTTGGAGTGCAG----------AAAAATCAACCGGCATTAGCA
AT3G60810.    AAAAAACCAGATTATCTAGGAGTACAA----------AAGAACGAG---AGATTGGCT
LOC_Os03g6    ACGAAGCCAAGGTACCTGGGCGTGCAG----------AAGAGTCCTCCATCGCTGGCT
Cre11.g468    CCTCGCCCGGGCAACCTGGGCATCATCGACTAC---GGCGCAGGCGTGCAGACCCTGAAC
Vocar20006    CCGCGTCCTAGCAATCTTGGCATTCAGGATTAC---GGCGCCGGTATCCAGACGCTGTCT
Pp1s223_13    GGCCGGACGTGGTTTCCGTGGGAGCAGACCACGCCGGAGCAATTTGAGCCAGAGCAAGGC

n101401_S.    CTGTGCCCTCCAACTCCGAATTGCATTTCAACCTCTGAA---GAAGCAAACGATCCATCT
Pp1s10_47V    CTGTGCCCACCAACTCCTGGTTGTATCTCCACCTCAGAG---GAGCTGAACGACCCCACC
LJFgene3      CTGTGCCCGGCAACGAAGAATTGCGTGTCAACCTCTGAG---AATATCAGTGATCGCACA
LJFgene1      CTGTGTCCGGCAACTAAGAACTGCGTGTCAACCTCTGAG---AATATCAGTGATCGCACA
LJFgene14     CTGTGTCCGCCAACTAAGAACTGCGTGTCAACCTCTGAG---AATATCAGCGATCGCACA
Phvul.010G    CTGTGTCCAGCAACAAAGAACTGTGTATCAACCTCTGAG---AATATCAGTGATCGCACG
LJFgene8      CTGTGTCCGGTAACTAGGAACTGCGTATCAACCTCTGAG---AATATCAGTGATCGCACT
Medtr4g023    CTGTGTCCAGCAACTAAGAATTGTGTATCAACCTCTGAG---AATGTCAATGATCTCACC
LJFgene9      CTATGTCCGGCAACTAAGAACTGCATATCGACATCTGAA---AATGTCACTAACCTCACA
AT3G60810.    CTATGTCCTGCCACTAATAACTGTATCTCCACTTCTGAG---AATATCAGTGATCGAGTG
LOC_Os03g6    CTGTGCCCTGCCACCAACAATTGCGTTTCCACTTGTGAG---GACATCACTGACTCCATT
Cre11.g468    CTGTGCCCTCCCTCGCCCAACTGCATTGCCACCTCCGAG---GAGGGCAACGACCGCACC
Vocar20006    CTCTGCCCGTCTACCCCCAACTGCATCGCGACATCGGAG---GAGGGCAATGACCGCGAG
Pp1s223_13    ACCTGCTCCACATGCATCGGCGTTGTGGATGACACTTTGGGGTCTTGCAGTGCCACCACC

n101401_S.    CACTACGTACCTCCGTGGACGTATAATCCAGAAGATGGACGTGGAAGAAAAAAC---CCC
Pp1s10_47V    CATTATGTTCCTCCATGGACGTACAATCCTCCAGATGGCCGTGGTCGGAAGAAC---CCT
LJFgene3      CATTATGCTCCTCCATGGAACTATAATCCT---------GAAGGTAGGAAAAAA---CCT
LJFgene1      CATTATGCTCCTCCATGGAACTATAATCCT---------GAAGGTAGGAAAAAA---CCT
LJFgene14     CATTATGCTCCTCCATGGAACTATAATCCT---------GAAGGAAGGAAAAAA---CCT
Phvul.010G    CATTATGCTCCTCCTTGGAACTATAATCCT---------GAAGGCAGGAAAAAT---CCT
LJFgene8      CATTATGCTCCTCTTTGGAACTACAATCCT---------GAAGGTAGGAAAAAC---CCT
Medtr4g023    CATTATGCTCCTCCTTGGAACTATAATCCT---------GAAGGCAGGAAAAGT---CCA
LJFgene9      CATTACACTCCTCCTTGGAACTACAATCCT---------GAAGGTAGGAAAGAT---CAT
AT3G60810.    CATTATGCTCCACCATGGAACTATAAT------------GGAGGAAGGAAAACA---CCT
LOC_Os03g6    CACTACGCTCCCCCATGGAACTATAACCCCAAG------GACGGACGCAGGGCTAAACCC
Cre11.g468    CACTATGCCCCTCCCCTGACCTACAACCCCGAGGATGGCCGCGGCAAGAAGGGC---CCG
Vocar20006    CACTACGCGCCCCCGCTCACGTACAACCCGCAGGATGGTCGTGGCCGCAAGAAC---CCC
Pp1s223_13    AACTGTGTTTCCTCCTTCGACGACAGGCCGAGCTTCTTCACAGCACCGTGGGAGTTTCCG

n101401_S.    GCCACCAGAGAGCAGGCCATGAAAGAGCTTCTCGCTGTTATATCTTCAACCAAG---CCG
Pp1s10_47V    GCTTCTAAAGAAAAAGCTATGGCGGAGCTCATTGATGCTATCAAGTCAACAAAG---CCG
LJFgene3      GTGAACAGAGAGGAAGCAATGGAGGAACTGATAGACGTGATAGAATCAACAACA---CCA
LJFgene1      GTGAGCAGGGAAGAAGCAATGGAGGAACTTATAGACGTGATAGAATCAACAACA---CCA
LJFgene14     GTGAGCAGAGAGGAAGCAATGGAGGAACTGATAGACGTGATCGAATCAACAACA---CCA
Phvul.010G    GTGAGCAAAGAGGAAGCCATGGAGGAACTGATAGACGTGATTGAATCAACAACA---CCA
LJFgene8      GTGAGCAGAGAAGAGGCAATGGAGGAACTGATAGACGTGATAGAATCAACAACA---CCA
Medtr4g023    GTGAGCAGAGAAGAGGCAGTGGAGGAACTAATAGAAGTGATAGAATTAACAAGA---CCA
LJFgene9      GTGAGCAAA---GAGGCAATGGAGGAACTGATAGATGTGATAGAATCGACAATACTACCA
AT3G60810.    GTGAACAGGCAAGTTGCAATGAAAGAGCTTCTTAATGTGATTAAGTCGGTGAAA---CCG
```

```
LOC_Os03g6    ATCACCAAACATGAAGCCATCAACCAGCTCATTCAAGTCGTCACACAAACAAAG---CCT
Cre11.g468    GCCAGCCAGGAGAAGGCCATGGGCGAGCTGGTGGAGGCGGTCAAGAAGCTCAAG---CCC
Vocar20006    GCTACTCAGGAACAGGCCATGGCCGAGTTGGTGTCTGTGGTCAAGACCCTTCAG---CCG
Pp1s223_13    GGATCTTTGAGGGCTGCTGTTAGTAATTTACAAGAAGCTTTAGAGGGTTCCGGA------

n101401_S.    GAAAACTATACTCCAAACATCGTCAAGAACACTGACGATTACGTCTACGTTGAGTATCAA
Pp1s10_47V    GACAACTTCACGCCTCGAATCGTTAAGCAGACTGATGATTACGTTTATGTCGAGTACTCA
LJFgene3      GACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAGTACCAA
LJFgene1      GACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAGTACCAA
LJFgene14     GACAAATTTTCACCACGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAGTACCAA
Phvul.010G    GACAAATTTACACCAAGGATAGTTGAAAGGAAAGAAGACTATATTCGTGTGGAGTATCAA
LJFgene8      GACAAATTTACACCACGAATAGTTGAAAGGAAGGAAGACTATATTCATGTGGAGTACCAA
Medtr4g023    GACAAATTTACACCAAAAATAGTTGAAAGGAAAGAAGACTATGTTCGAGTTGAGTATCGA
LJFgene9      GAAAATTTTACACCAAGGATTGTAGAAAGAACAGAAGATTATCTTAGATTGGAATACCAA
AT3G60810.    GATAAATTTACACCGCGGATTGTGGAGAAGAAGGATGACTACGTTCATGTGGAATATGAA
LOC_Os03g6    GACAACTTCACTCCTCGCCTAGTAGAGAAAACAGATGACTATGTTCGAGTTGAATACGAG
Cre11.g468    GACGGCTTCACCCCCAAGATCATCAAGCAGACCGACGACTACCTGTACGTCGAGTACGAG
Vocar20006    GACGGATTCACCCCCAAGATCATCAAGCAGACCCCGAACTACTTGTACGTGGAGTACGAG
Pp1s223_13    ------------GCTTTCATCAAGGAGAAGAGTGACAGATACATTTACGCAGTTTTCAAA

n101401_S.    TCTCCCTATCTGGGTTTTGTAGACGATGTCGAGTTTTGGTTCCCG---CCTGGGAATCGA
Pp1s10_47V    AGCCCTCTTGTTGGCTTTGTGGACGATGTCGAATTTTGGTTTCCT---CCTGGCAATCGG
LJFgene3      AGCTCAATTTTGGGGTTTGTAGATGATGTTGAGTTCTGGTTCCCA---CCGGGTAAGGGT
LJFgene1      AGCTCAATCTTGGGGTTTGTGGATGATGTTGAGTTCTGGTTTCCT---CCGGGTAAGGGT
LJFgene14     AGCTCAATCTTGGGGTTTGTGGATGATGTTGAGTTCTGGTTTCCA---CCCGGTAAGGGT
Phvul.010G    AGCTCAATCTTGGGGTTTGTGGATGATGTTGAGTTCTGGTTTCCA---CCTAATAAGGGT
LJFgene8      AGCTCAATCTTGGGGTTTGTGCATGATGTTGAGTTCTGGTTTCCA---CTGGGTAAGGGT
Medtr4g023    AGCTCAATCTTGGGGTTTGTAGACGATGTTGAGTTTTGGTTCCCA---CCAGGTAAGGGC
LJFgene9      AGT---------------GTATACAAGCCACAAATTTTAACTTCA---ATGTCACCAATA
AT3G60810.    AGTCCAATCTTGGGGTTAGTAGACGATGTTGAATTCTTGTTCACT---CCTGGGAAGAAC
LOC_Os03g6    AGTCCTATATTCGGGTTTGTAGATGATGTGGAGTTCTGGTTCCCT---CCTGGTAACAAA
Cre11.g468    AGCCCGCTCATGGGGTTCATTGACGATGTGGAGTTCTGGTTCAAG---CCCGGCCCCGGC
Vocar20006    AGCCCCATTATGGGGTTCATCGACGATGTGGAGTTCTGGTTCAAG---CCAGGTCCCGGC
Pp1s223_13    GGAGAT---GATGGCGTTAGCGACGATGTGGAATTCTTATTCAGTGATCCCTCTGTTGAT

n101401_S.    TCGCTGGTTGAGTACAGGTCGGCTTCTCGCGTGGGATCGTCG---GATTTCGATGCTAAT
Pp1s10_47V    TCGTTGGTAGAGTATCGATCTGCATCTCGGAGAGATGCCATA---GACTTTGGCTTCAAC
LJFgene3      TCTACTGTGGAGTACCGATCTGCATCTCGGTTAGGAAACTTT---GATTTTGATGTGAAC
LJFgene1      TCTACTGTGGAGTATCGTTCTGCATCTCGGTTGGGAAACTTT---GATTTTGATGTGAAC
LJFgene14     TCTACTGTGGAGTATCGATCTGCATCTCGGTTGGGAAACTTT---GATTTTGATGTGAAC
Phvul.010G    TCTACTGTGGAGTATCGATCAGCATCTCGGCTGGGGAACTTT---GATTTTGATCTGAAT
LJFgene8      TCTACTGTGGAGTATCGATCTGCATCTCGGTTGGGGAACTTT---GATTTTGATGTGAAT
Medtr4g023    TCAATCGTGGAGTATCGATCTGCGTCAAGGTTGGGAAACTTT---GATTTTGATGTGAAT
LJFgene9      TCATTGTATGCAGAAAAAATGAATAGTAACTTTTTACTATTA---GAC------------
AT3G60810.    TCGAAAGTGGAATATCGATCTGCATCCCGTAAAGGGAACTTC---GACTTTGATGTCAAT
LOC_Os03g6    TCCATTGTTCAGTACCGATCAGCATCTCGATCAGGATTCATT---GACTTCAACGCCAAC
Cre11.g468    AGCCGCGTGGAGTACCGCAGCGCCAGCCGCGTGGGCGAGTCG---GACGGCAACATCAAC
Vocar20006    GCCCGAGTGGAGTACCGCTCCGCGTCACGCGTTGGGGAGTCT---GACGGCAACATCAAC
Pp1s223_13    GCAACTGTAAATGTGAGGTCTGCCTCTCGTGCTAAAGATTACAAAGACAGCGGCAGAAAT

n101401_S.    CGCAAGAGAATCAGAGCTTTGCGAAAGGCGCTGGAGAAGAAAGGCTGGCAATCGATTGGC
Pp1s10_47V    CGCAAGCGAATTAAGGCTCTGCGTCAAGCTCTGGAGCGCTATGGATGGGAATCTATTGGA
LJFgene3      AGAAAAAGAATAAAGGCACTGCGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCAAGAC
LJFgene1      AGAAAAAGAATAAAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGGCATCTCAAGAC
LJFgene14     AGAAAAAGAATAAAGGCACTGAGACAAGAGTTGGAGAAGAAAGGATGGACATCTCAAGAT
Phvul.010G    AGAAAAAGAATAAAGGCACTGCGACAAGAGCTGGAGAAGAAAGGATGGGCATCTGAAGAC
LJFgene8      AAGAAAAGAATAAAGGTATGT---------TTGTATCATTCCTTTGTGCTGTCTCGG---
Medtr4g023    AGAAAAAGAATAAAGGCATTGCGACAAGAGTTAGAGAAGAAAGGATGGGCATCTCAAGAT
LJFgene9      ------------------------------------------------------------
AT3G60810.    AGAAAGCGAATAAAGGCATTGCGACAAGAGCTAGAGAAGAAGGGATGGGTATCAGAGAAC
LOC_Os03g6    AAGAAAAGAGTCAAGGCGCTCAGGTTGGCACTGGAAAACAAAGGCTGGGCTTCAGAAAGC
Cre11.g468    CGCAAGCGCATCAAGGCCATCCGCCAGGAGCTGGAGACCAAGGGCTGGCGCAGCACGGGC
Vocar20006    CGCAAGCGCATCCGGGCCATCCGCCAGGAGCTCGAGAAGGAGGGTTGGCGCAGCACCGGC
Pp1s223_13    CGCAAAAGACTCGAAGCCCTACGCATGGAGCTTGGCTGGGAACAGGTGCCAATTTTGAGG
```

```
n101401_S.    TTC-------------------------------------------------------
Pp1s10_47V    TTC-------------------------------------------------------
LJFgene3      ACCATA----------------------------------------------------
LJFgene1      ACCATA----------------------------------------------------
LJFgene14     ACCATA----------------------------------------------------
Phvul.010G    ACAATT----------------------------------------------------
LJFgene8      ------------------------------------------------------------
Medtr4g023    ACTACTATA-------------------------------------------------
LJFgene9      ------------------------------------------------------------
AT3G60810.    AGCTTC----------------------------------------------------
LOC_Os03g6    ACCATC----------------------------------------------------
Cre11.g468    TTC-------------------------------------------------------
Vocar20006    TTC-------------------------------------------------------
Pp1s223_13    AACCGGCAGAGGCGGCTCTTTTTCATTGAATCTCCGTGGGACACTTTCGGTCCGGAGCCA

n101401_S.    ------------------------------------------
Pp1s10_47V    ------------------------------------------
LJFgene3      ------------------------------------------
LJFgene1      ------------------------------------------
LJFgene14     ------------------------------------------
Phvul.010G    ------------------------------------------
LJFgene8      ------------------------------------------
Medtr4g023    ------------------------------------------
LJFgene9      ------------------------------------------
AT3G60810.    ------------------------------------------
LOC_Os03g6    ------------------------------------------
Cre11.g468    ------------------------------------------
Vocar20006    ------------------------------------------
Pp1s223_13    CCTCCTACGATTGACTATAAAAATGGAATAGACTTTGTGCCGGAC
```

APPENDIX L.

DIVERSE PLANT FAMILY FULL SYNONYMOUS/NONSYNONYMOUS DATA TABLE

Ortholog SNAP analysis											
Compare	**vs**	**Sd**	**Sn**	**S**	**N**	**ps**	**pn**	**ds**	**dn**	**ds/dn**	**ps/pn**
Smo	Ppa47v6	94.67	55.33	110.00	352.00	0.86	0.16	NA	0.00	NA	5.47
Smo	LJFgene3	85.00	73.00	103.00	350.00	0.83	0.21	NA	0.00	NA	3.96
Smo	LJFgene1	90.00	74.00	102.33	350.67	0.88	0.21	NA	0.00	NA	4.17
Smo	LJFgene14	90.00	71.00	102.33	350.67	0.88	0.20	NA	0.00	NA	4.34
Smo	Pvu	88.83	74.17	102.67	350.33	0.87	0.21	NA	0.00	NA	4.09
Smo	LJFgene8	89.00	85.00	99.83	338.17	0.89	0.25	NA	0.00	NA	3.55
Smo	Mtr	84.00	73.00	102.50	350.50	0.82	0.21	NA	0.00	NA	3.93
Smo	LJFgene9	68.83	113.17	81.33	281.67	0.85	0.40	NA	0.00	NA	2.11
Smo	Ath	86.83	82.17	100.50	346.50	0.86	0.24	NA	0.00	NA	3.64
Smo	Osa	88.50	87.50	105.83	353.17	0.84	0.25	NA	0.00	NA	3.38
Smo	Cre	86.33	95.67	107.50	354.50	0.80	0.27	NA	0.00	NA	2.98
Smo	Vca	92.67	89.33	108.00	354.00	0.86	0.25	NA	0.00	NA	3.40
Smo	Ppa13v6	85.17	211.83	105.33	356.67	0.81	0.59	NA	0.00	NA	1.36
Ppa47v6	LJFgene3	117.33	185.67	157.83	529.17	0.74	0.35	3.55	0.47	7.50	2.12
Ppa47v6	LJFgene1	110.17	178.83	150.50	509.50	0.73	0.35	2.80	0.47	5.91	2.09
Ppa47v6	LJFgene14	82.17	95.83	110.50	375.50	0.74	0.26	3.57	0.31	11.45	2.91
Ppa47v6	Pvu	115.50	180.50	151.33	508.67	0.76	0.35	NA	0.00	NA	2.15
Ppa47v6	LJFgene8	112.17	185.83	147.00	492.00	0.76	0.38	NA	0.00	NA	2.02
Ppa47v6	Mtr	117.17	175.83	149.00	505.00	0.79	0.35	NA	0.00	NA	2.26
Ppa47v6	LJFgene9	70.67	122.33	88.00	302.00	0.80	0.41	NA	0.00	NA	1.98
Ppa47v6	Ath	115.50	177.50	146.17	492.83	0.79	0.36	NA	0.00	NA	2.19
Ppa47v6	Osa	122.00	178.00	151.83	484.17	0.80	0.37	NA	0.00	NA	2.19
Ppa47v6	Cre	129.67	205.33	158.33	504.67	0.82	0.41	NA	0.00	NA	2.01
Ppa47v6	Vca	144.83	192.17	159.00	510.00	0.91	0.38	NA	0.00	NA	2.42
Ppa47v6	Ppa13v6	112.00	325.00	155.83	513.17	0.72	0.63	2.38	1.40	1.71	1.13
LJFgene3	LJFgene1	18.50	9.50	142.00	512.00	0.13	0.02	0.14	0.02	7.62	7.02
LJFgene3	LJFgene14	19.00	23.00	106.67	382.33	0.18	0.06	0.20	0.06	3.24	2.96
LJFgene3	Pvu	40.83	46.17	142.83	511.17	0.29	0.09	0.36	0.10	3.74	3.17
LJFgene3	LJFgene8	31.50	49.50	136.17	487.83	0.23	0.10	0.28	0.11	2.54	2.28
LJFgene3	Mtr	56.50	61.50	141.83	509.17	0.40	0.12	0.57	0.13	4.31	3.30
LJFgene3	LJFgene9	47.00	92.00	85.33	304.67	0.55	0.30	0.99	0.39	2.57	1.82
LJFgene3	Ath	94.83	117.17	136.33	490.67	0.70	0.24	1.97	0.29	6.85	2.91
LJFgene3	Osa	111.67	126.33	142.17	481.83	0.79	0.26	NA	0.00	NA	3.00
LJFgene3	Cre	127.00	205.00	145.00	494.00	0.88	0.42	NA	0.00	NA	2.11
LJFgene3	Vca	131.17	205.83	146.50	498.50	0.90	0.41	NA	0.00	NA	2.17
LJFgene3	Ppa13v6	119.67	314.33	144.50	503.50	0.83	0.62	NA	0.00	NA	1.33
LJFgene1	LJFgene14	15.00	23.00	105.83	383.17	0.14	0.06	0.16	0.06	2.51	2.36
LJFgene1	Pvu	38.33	40.67	144.50	518.50	0.27	0.08	0.33	0.08	3.95	3.38
LJFgene1	LJFgene8	25.83	49.17	137.83	495.17	0.19	0.10	0.22	0.11	2.02	1.89

LJFgene1	Mtr	56.00	57.00	141.83	512.17	0.39	0.11	0.56	0.12	4.65	3.55
LJFgene1	LJFgene9	49.17	89.83	84.83	305.17	0.58	0.29	1.11	0.37	2.97	1.97
LJFgene1	Ath	91.33	119.67	138.00	498.00	0.66	0.24	1.61	0.29	5.54	2.75
LJFgene1	Osa	116.83	128.17	144.17	488.83	0.81	0.26	NA	0.00	NA	3.09
LJFgene1	Cre	130.50	209.50	147.00	501.00	0.89	0.42	NA	0.00	NA	2.12
LJFgene1	Vca	133.00	213.00	148.33	505.67	0.90	0.42	NA	0.00	NA	2.13
LJFgene1	Ppa13v6	123.83	321.17	146.50	510.50	0.85	0.63	NA	0.00	NA	1.34
LJFgene14	Pvu	29.00	30.00	106.50	382.50	0.27	0.08	0.34	0.08	4.08	3.47
LJFgene14	LJFgene8	26.00	51.00	102.50	368.50	0.25	0.14	0.31	0.15	2.02	1.83
LJFgene14	Mtr	43.00	40.00	106.17	382.83	0.41	0.10	0.58	0.11	5.18	3.88
LJFgene14	LJFgene9	48.83	107.17	84.00	306.00	0.58	0.35	1.12	0.47	2.37	1.66
LJFgene14	Ath	74.67	65.33	103.67	379.33	0.72	0.17	2.42	0.20	12.37	4.18
LJFgene14	Osa	91.17	86.83	108.50	380.50	0.84	0.23	NA	0.00	NA	3.68
LJFgene14	Cre	99.33	130.67	109.33	376.67	0.91	0.35	NA	0.00	NA	2.62
LJFgene14	Vca	107.50	130.50	109.50	376.50	0.98	0.35	NA	0.00	NA	2.83
LJFgene14	Ppa13v6	86.50	231.50	107.33	381.67	0.81	0.61	NA	0.00	NA	1.33
Pvu	LJFgene8	41.50	69.50	138.33	494.67	0.30	0.14	0.38	0.16	2.46	2.14
Pvu	Mtr	57.67	70.33	142.67	511.33	0.40	0.14	0.58	0.15	3.82	2.94
Pvu	LJFgene9	46.67	85.33	84.83	305.17	0.55	0.28	0.99	0.35	2.83	1.97
Pvu	Ath	90.67	117.33	138.33	497.67	0.66	0.24	1.55	0.28	5.49	2.78
Pvu	Osa	122.17	134.83	145.00	488.00	0.84	0.28	NA	0.00	NA	3.05
Pvu	Cre	126.50	206.50	147.83	500.17	0.86	0.41	NA	0.00	NA	2.07
Pvu	Vca	139.67	203.33	149.17	504.83	0.94	0.40	NA	0.00	NA	2.32
Pvu	Ppa13v6	120.83	321.17	147.33	509.67	0.82	0.63	NA	0.00	NA	1.30
LJFgene8	Mtr	55.67	82.33	136.17	490.83	0.41	0.17	0.59	0.19	3.11	2.44
LJFgene8	LJFgene9	48.50	91.50	84.67	305.33	0.57	0.30	1.08	0.38	2.83	1.91
LJFgene8	Ath	95.50	138.50	136.33	484.67	0.70	0.29	2.04	0.36	5.67	2.45
LJFgene8	Osa	114.00	142.00	140.50	471.50	0.81	0.30	NA	0.00	NA	2.69
LJFgene8	Cre	129.17	225.83	145.50	490.50	0.89	0.46	NA	0.00	NA	1.93
LJFgene8	Vca	126.33	221.67	145.67	493.33	0.87	0.45	NA	0.00	NA	1.93
LJFgene8	Ppa13v6	117.17	316.83	143.00	496.00	0.82	0.64	NA	0.00	NA	1.28
Mtr	LJFgene9	52.83	89.17	85.00	305.00	0.62	0.29	1.32	0.37	3.57	2.13
Mtr	Ath	96.50	126.50	136.00	494.00	0.71	0.26	2.19	0.31	6.99	2.77
Mtr	Osa	125.83	134.17	143.33	483.67	0.88	0.28	NA	0.00	NA	3.16
Mtr	Cre	132.83	211.17	145.67	496.33	0.91	0.43	NA	0.00	NA	2.14
Mtr	Vca	131.67	214.33	146.83	501.17	0.90	0.43	NA	0.00	NA	2.10
Mtr	Ppa13v6	122.83	317.17	145.50	508.50	0.84	0.62	NA	0.00	NA	1.35
LJFgene9	Ath	62.00	107.00	82.33	301.67	0.75	0.35	NA	0.00	NA	2.12
LJFgene9	Osa	65.67	113.33	86.67	303.33	0.76	0.37	NA	0.00	NA	2.03
LJFgene9	Cre	85.83	154.17	88.50	301.50	0.97	0.51	NA	0.00	NA	1.90
LJFgene9	Vca	80.33	148.67	89.33	300.67	0.90	0.49	NA	0.00	NA	1.82

LJFgene9	Ppa13v6	65.50	220.50	85.33	304.67	0.77	0.72	NA	0.00	NA	1.06
Ath	Osa	121.00	138.00	140.33	474.67	0.86	0.29	NA	0.00	NA	2.97
Ath	Cre	126.17	216.83	145.33	490.67	0.87	0.44	NA	0.00	NA	1.96
Ath	Vca	126.50	205.50	146.00	493.00	0.87	0.42	NA	0.00	NA	2.08
Ath	Ppa13v6	118.33	317.67	143.17	498.83	0.83	0.64	NA	0.00	NA	1.30
Osa	Cre	105.50	208.50	152.33	483.67	0.69	0.43	1.93	0.64	3.00	1.61
Osa	Vca	111.17	204.83	152.17	483.83	0.73	0.42	2.74	0.62	4.39	1.73
Osa	Ppa13v6	123.67	299.33	149.83	489.17	0.83	0.61	NA	0.00	NA	1.35
Cre	Vca	88.17	104.83	159.50	515.50	0.55	0.20	1.00	0.24	4.22	2.72
Cre	Ppa13v6	114.50	312.50	156.50	518.50	0.73	0.60	2.78	1.22	2.28	1.21
Vca	Ppa13v6	122.17	313.83	157.00	524.00	0.78	0.60	NA	0.00	NA	1.30

BIBLIOGRAPHY

[1] The Biology of *Glycine max* (L.) Merr. (Soybean) by Canadian Food Inspection Agency.

[2] *Glycine max* (L.) Merr., USDA Plants Profile, http://plants.usda.gov/java/profile?symbol=GLMA4, July 2013.

[3] Chan C, Qi X, Li M-, Wong F-, Lam H-. Recent developments of genomic research in soybean. Journal of Genetics and Genomics 2012;39(7):317-24.

[4] Hymowitz T. On the domestication of the soybean. Econ Bot 1970;24(4):408-21.

[5] Hymowitz T, Harlan JR. Introduction of soybean to North America by Samuel Bowen in 1765. Econ Bot 1983;37(4):371-9.

[6] Soystats, American Soybean Association, 2011. http://www.soystats.com/2011/Default-frames.htm.

[7] USDA NASS at http://www.nass.usda.gov/QuickStats/index2.jsp. May 2013.

[8] DOE JGI press release, DOE JGI Releases Soybean Genome Assembly. Jan 2008. http://www.jgi.doe.gov/News/news_1_17_08.html

[9] Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, Gill N, Joshi T, Libault M, Sethuraman A, Zhang X-, Shinozaki K, Nguyen HT, Wing RA, Cregan P, Specht J, Grimwood J, Rokhsar D, Stacey G, Shoemaker RC, Jackson SA. Genome sequence of the palaeopolyploid soybean. Nature 2010;463(7278):178-83.

[10] Hurles M. Gene duplication: The genomic trade in spare parts. PLoS Biology 2004;2(7).

[11] Young ND, Bharti AK. Genome-enabled insights into legume biology. Annu Rev Plant Biol 2012: 63, 283-305.

[12] Thornton JW, DeSalle R. Gene family evolution and homology: Genomics meets phylogenetics. Annual Review of Genomics and Human Genetics 2000;1(2000):41-73.

[13] Hartwell, Leland H., Leroy Hood, Michael L. Goldberg, Ann E. Reynolds, and Lee M. Silver. *Genetics: From Genes to Genomes*. 4th ed. New York: McGraw-Hill, 2011, p. 345,696. Print.

[14] Shoemaker RC, Schlueter J, Doyle JJ. Paleopolyploidy and gene duplication in soybean and other legumes. Curr Opin Plant Biol 2006;9(2):104-9.

[15] Wessler, Susan R., Sean B. Carroll, and John Doebley. "Large-Scale Chromosomal Changes." *Introduction to Genetic Analysis*. By Anthony J. F. Griffiths. 10th ed. New York: W.H. Freeman and, 2012. 592-99;290-91. Print.

[16] Micklos, David A., Uwe Hilgert, and Bruce Nash. *Genome Science: A Practical and Conceptual Introduction to Molecular Genetic Analysis in Eukaryotes*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 2013, p. 24-25. Print.

[17] Science Primer at NCBI at http://www.ncbi.nlm.nih.gov/About/primer/est.html. May 2013.

[18] Nelson RT, Shoemaker R. Identification and analysis of gene families from the duplicated genome of soybean using EST sequences. BMC Genomics 2006;7.

[19] dbEST at NCBI: http://www.ncbi.nlm.nih.gov/nucest. May 2013.

[20] Ho ES, Gunderson SI, Duffy S. A multispecies polyadenylation site model.BMC Bioinformatics. 2013;14 Suppl 2:S9. doi: 10.1186/1471-2105-14-S2-S9. Epub 2013 Jan 21.

[21] Joshi CP. Putative polyadenylation signals in nuclear genes of higher plants: A compilation and analysis. Nucleic Acids Res 1987;15(23):9627-40.

[22] Ji, G., Zhang, H., Wu, X., Tang, M.Identification of plant messenger RNA polyadenylation sites using length-variable second order markov model. Conference proceedings - IEEE international conference on systems, man and cybernetics; 2011. p. 920-924.

[23] Sherstnev A, Duc C, Cole C, Zacharaki V, Hornyik C, Ozsolak F, Milos PM, Barton GJ, Simpson GG. Direct sequencing of arabidopsis thaliana RNA reveals patterns of cleavage and polyadenylation. Nature Structural and Molecular Biology 2012;19(8):845-52.

[24] Wu X, Chen L, Ji G. Prediction of plant poly(A) sites based on GHMM-RWT. Information 2012;15(4):1809-21.

[25] Labadorf A, Link A, Rogers MF, Thomas J, Reddy ASN, Ben-Hur A. Genome-wide analysis of alternative splicing in Chlamydomonas reinhardtii. BMC Genomics 2010, 11:114.

[26] Reddy AS, Rogers MF, Richardson DN, Hamilton M, Ben-Hur A. Deciphering the plant splicing code: experimental and computational approaches for predicting alternative

splicing and splicing regulatory elements. Front Plant Sci. 2012;3:18. doi: 10.3389/fpls.2012.00018. Epub 2012 Feb 7.

[27] Graveley BR, Hertel KJ, Maniatis TOM. The role of U2AF35 and U2AF65 in enhancer-dependent splicing. RNA 2001;7(6):806-18.

[28] Simpson CG, Jennings SN, Clark GP, Thow G, Brown JWS. Dual functionality of a plant U-rich intronic sequence element. Plant Journal 2004;37(1):82-91.

[29] Goodall GJ, Filipowicz W. Different effects of intron nucleotide composition and secondary structure on pre-mesenger-RNA splicing in monocot and dicot plants. EMBO J. 1991: 10, p. 2635-2644.

[30] Brown JWS. A catalogue of splice junction and putative branch point sequences from plant introns. Nucleic Acid Res. 1986: 14, p. 9549-9559.

[31] Simpson CG, Thow G, Clark GP, Jennings SN, Watters JA, Brown JWS. Mutational analysis of a plant branchpoint and polypyrimidine tract required for constitutive splicing of a mini-exon. RNA 2002;8(1):47-56.

[32] http://www.genome.jp/kegg/catalog/codes1.html. May 2013.

[33] Thomas J, Palusa SG, Prasad KVSK, Ali GS, Surabhi G-, Ben-Hur A, Abdel-Ghany SE, Reddy ASN. Identification of an intronic splicing regulatory element involved in auto-regulation of alternative splicing of SCL33 pre-mRNA. Plant Journal 2012;72(6):935-46.

[34] Pertea M, Mount SM, Salzberg SL. A computational survey of candidate exonic splicing enhancer motifs in the model plant arabidopsis thaliana. BMC Bioinformatics 2007;8.

[35] Kandoth C, Ercal F, Frank RL. A framework for automated enrichment of functionally significant inverted repeats in whole genomes. BMC Bioinformatics 2010;11(SUPPL. 6).

[36] Turner M, Yu O, Subramanian S. Genome organization and characteristics of soybean microRNAs. BMC Genomics 2012;13(1).

[37] Wang Y, Li P, Cao X, Wang X, Zhang A, Li X. Identification and expression analysis of miRNAs from nitrogen-fixing soybean nodules. Biochem Biophys Res Commun 2009;378(4):799-803.

[38] Shen W, Chen M, Wei G, Li Y (2012) MicroRNA Prediction Using a Fixed-Order Markov Model Based on the Secondary Structure Pattern. PLoS ONE 7(10): e48236. doi:10.1371/journal.pone.0048236.

[39] The Pfam protein families database: M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E.L.L. Sonnhammer, S.R. Eddy, A. Bateman, R.D. FinnNucleic Acids Research (2012) Database Issue 40:D290-D301.

[40] http://www.pantherdb.org/about.jsp.

[41] PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Huaiyu Mi, Anushya Muruganujan and Paul D. Thomas Nucl. Acids Res. (2012) doi: 10.1093/nar/gks1118.

[42] Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Smirnov S, Nikolskaya AN, Rao BS, Mekhedov SL, Sverlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. The COG database: An updated vesion includes eukaryotes. BMC Bioinformatics 2003;4.

[43] http://genome.jgi.doe.gov/Tutorial/tutorial/kog.html. May 2013.

[44] Berman H, Henrick K, Nakamura H. Announcing the worldwide protein data bank. Nat Struct Biol 2003;10(12):980.

[45] http://www.phytozome.net/. Accessed Aug 7, 2013.

[46] NCBI at a Glance: Our Mission. http://www.ncbi.nlm.nih.gov/About/glance/ourmission.html. Aug 2013.

[47] NCBI at a Glance: Programs and Activities. http://www.ncbi.nlm.nih.gov/About/glance/programs.html.

[48] About – DNA Subway. http://dnasubway.iplantcollaborative.org/about/.

[49] http://dnasubway.iplantcollaborative.org/. May 2013.

[50] Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.-F., Guindon S., Lefort V., Lescot M., Claverie J.-M., Gascuel O. *Phylogeny.fr: robust phylogenetic analysis for the non-specialist* Nucleic Acids Research. 2008 Jul 1; 36 (Web Server Issue):W465-9. Epub 2008 Apr 19. (PubMed).

[51] http://www.expasy.org/features. May 2013.

[52] Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31(13):3784-8.

[53] http://meme.nbcr.net/meme/. May 2013.

[54] Timothy L. Bailey and Charles Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California, 1994.

[55] About GenomeNet. http://www.genome.jp/en/about.html. May 2013.

[56]GenomeNet: Bioinformatics tools. http://www.genome.jp/en/gn_tools.html. May 2013.

[57] Mikita Suyama, David Torrents, and Peer Bork (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34, W609-W612.

[58] www.hiv.lanl.gov. May 2013.

[59] Korber B. (2000). HIV Signature and Sequence Variation Analysis. Computational Analysis of HIV Molecular Sequences, Chapter 4, pages 55-72. Allen G. Rodrigo and Gerald H. Learn, eds. Dordrecht, Netherlands: Kluwer Academic Publishers.

[60] Higo, K., Y. Ugawa, M. Iwamoto and T. Korenaga (1999) Plant cis-acting regulatory DNA elements (PLACE) database:1999. Nucleic Acids Research Vol.27 No.1 pp. 297-300.

[61] Prestridge, D.S. (1991) SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements. CABIOS 7, 203-206.

[62] Yu CS, Lin CJ, Hwang JK: Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Science 2004, 13:1402-1406.

[63] Yu CS, Chen YC, Lu CH, Hwang JK: Prediction of protein subcellular localization. Proteins: Structure, Function and Bioinformatics 2006, (in press).

[64] Yu CS, Lin CJ, Hwang JK: Prediction of Subcellular Locations by Support Vector Machines Using Multiple Feature Vectors Based on n-peptide Compositions. (unpublished data).

[65] Jones DT. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292: 195-202.

[66] http://bioinf.cs.ucl.ac.uk/psipred/. May 2013.

[67] McGuffin LJ & Jones DT. (2003) Improvement of the GenTHREADER method for genomic fold recognition. Bioinformatics, 19, 874-881.

[68] CATH Documentation. http://www.cathdb.info/wiki/doku/?id=faq. May 2013.

[69] Lobley, A., Sadowski, M.I. & Jones, D.T. (2009) pGenTHREADER and pDomTHREADER: New Methods For Improved Protein Fold Recognition and Superfamily Discrimination. Bioinformatics. 25, 1761-1767.

[70] Overview of Prediction Methods. http://bioinf.cs.ucl.ac.uk/index.php?id=779.

[71] Background on Nucleic Acid dot Plots. http://www.vivo.colostate.edu/molkit/dnadot/bkg.html. May 2013.

[72] Roy A, Kucukural A, Zhang Y. I-TASSER: A unified platform for automated protein structure and function prediction. Nature Protocols 2010;5(4):725-38.

[73] Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9.

[74] http://zhanglab.ccmb.med.umich.edu/I-TASSER/. May 2013.

[75] Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A.; *Protein Identification and Analysis Tools on the ExPASy Server;* (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press (2005). pp. 571-607.

[76] PDB ID: 1V5S. N. Tochio, S. Koshiba, M. Inoue, T. Kigawa, S. Yokoyama, RIKEN Structural Genomics/Proteomics Initiative. Solution structure of kinase associated domain 1 of mouse MAP/microtubule affinity-regulating kinase 3.

[77] PDB ID: 1UP8. E. Garcia-Rodriguez, T. Ohshiro, T. Aibara, Y. Izumi, J. Littlechild. (2005) Enhancing effect of calcium and vanadium ions on thermal stability of bromoperoxidase from Corallina pilulifera. J Biol Inorg10(3):275-82.

[78] PDB ID: 1m40. G. Minasov, X. Wang, B.K. Shoichet (2002). An ultrahigh resolution structure of TEM-1 beta-lactamase suggests a role for Glu166 as the general base in acylation. J Am Chem Soc. 124(19):5333-40.

[79] PDB ID: 1GA0. G.V. Crichlow, M. Nukaga, V.R. Doppalapudi, J.D. Buynak, J.R. Knox (2001). Inhibition of class C beta-lactamases: structure of a reaction intermediate with a cephem sulfone. Biochemistry 40(21):6233-9.

[80] Majiduddin FK, Materon IC, Palzkill TG. Molecular analysis of beta-lactamase structure and function. International Journal of Medical Microbiology 2002;292(2):127-37.

[81] Pradel N, Delmas J, Wu LF, Santini CL, Bonnet R. Sec- and tat-dependent translocation of β-lactamases across the escherichia coli inner membrane. Antimicrob

Agents Chemother 2009;53(1):242-8.

[82] Li HY, Zhu YM, Chen Q, Conner RL, Ding XD, Li J, Zhang BB. Production of transgenic soybean plants with two anti-fungal protein genes via agrobacterium and particle bombardment. Biol Plant 2004;48(3):367-74.

[83] Ng TB, Ye XJ, Wong JH, Fang EF, Chan YS, Pan W, Ye XY, Sze SCW, Zhang KY, Liu F, Wang HX. Glyceollin, a soybean phytoalexin with medicinal properties. Appl Microbiol Biotechnol 2011;90(1):59-68.

[84] PDB ID: 1ZG3. Liu C-, Deavours BE, Richard SB, Ferrer J-, Blount JW, Huhman D, Dixon RA, Noel JP. Structural basis for dual functionality of isoflavonoid O-methyltransferases in the evolution of plant defense responses. Plant Cell 2006;18(12):3656-69.

VITA

Lisa Snoderly-Foster was born in 1981 in Springfield, Missouri as Lisa Snoderly to Gregory and Barbara Snoderly of Richland, Missouri. She attended Laquey High School in Laquey, Missouri and matriculated to Southwest Baptist University in Bolivar, Missouri in 1999. She graduated SBU in 2003 with a Bachelor of Science in Biology.

After an extended period in the workforce, Lisa reentered the educational arena in 2010 in pursuit of a teaching certification from Missouri University of Science and Technology. She graduated May of 2012 with a Bachelor of Arts degree in Biology and a certification in secondary education.

Lisa was accepted into the Applied and Environmental Biology graduate program at Missouri University of Science and Technology in 2012 and earned a Master of Science degree in 2014.